2018

Modelling and forecasting stock price movements with serially dependent determinants

Rasika Yatigammana

Shelton Peiris

Richard Gerlach

David Edmund Allen
Edith Cowan UNiversity

10.3390/risks6020052
Yatigammana, R., Peiris, S., Gerlach, R., & Allen, D. E. (2018). Modelling and forecasting stock price movements with serially dependent determinants. *Risks*, *6*(2), 52.

risks

MDPI

Article

Modelling and Forecasting Stock Price Movements with Serially Dependent Determinants

Rasika Yatigammana [1], Shelton Peiris [2,3], Richard Gerlach [4] and David Edmund Allen [2,5,6,*]

1 Central Bank of Sri-Lanka, Colombo 01, Sri Lanka; yatigammanar@hotmail.com
2 School of Mathematics and Statistics, The University of Sydney, Sydney 2006, Australia; shelton.peiris@sydney.edu.au
3 Department of Statistics, The University of Colombo, Colombo 03, Sri-Lanka
4 Discipline of Business Analytics, The University of Sydney, Sydney 2006, Australia; richard.gerlach@sydney.edu.au
5 Department of Finance, Asia University, Taichung 41354, Taiwan
6 School of Business and Law, Edith Cowan University, Joondalup 6027, Australia
* Correspondence: profallen2007@gmail.com

Received: 16 March 2018; Accepted: 1 May 2018; Published: 7 May 2018

Abstract: The direction of price movements are analysed under an ordered probit framework, recognising the importance of accounting for discreteness in price changes. By extending the work of Hausman et al. (1972) and Yang and Parwada (2012),This paper focuses on improving the forecast performance of the model while infusing a more practical perspective by enhancing flexibility. This is achieved by extending the existing framework to generate short term multi period ahead forecasts for better decision making, whilst considering the serial dependence structure. This approach enhances the flexibility and adaptability of the model to future price changes, particularly targeting risk minimisation. Empirical evidence is provided, based on seven stocks listed on the Australian Securities Exchange (ASX). The prediction success varies between 78 and 91 per cent for in-sample and out-of-sample forecasts for both the short term and long term.

Keywords: ordered probit; stock prices; auto-regressive; multi-step ahead forecasts

1. Introduction

There has been a significant growth in market micro-structure research, which is concerned with the study of the underlying process that translates the latent demands of investors into transaction prices and volumes (Madhavan 2000). The study of the time series properties of security prices has been central to market micro-structure research for many years. Madhavan (2000) asserts that frictions and departures from symmetric information do affect the trading process. Furthermore, insights into future price trends provides additional information useful in strategy formulation. As per financial economic theory, the asset returns cannot be easily predicted by employing statistical or other techniques and incorporating publicly available information. Nevertheless, recent literature bears evidence of successful forecasting of asset return signs; see for example, Breen et al. (1989); Leung et al. (2000); White (2000); Pesaran and Timmermann (2004) and Cheung et al. (2005). While having mean independence, it is statistically probable to have sign and volatility dependence in asset returns (Christoffersen and Diebold 2006).

The knowledge of the future direction of the stock price movement provides valuable guidance in developing profitable trading strategies. However, there is no clear consensus on the stochastic behaviour of prices or on the major factors determining the change in prices. In this context, theories of information asymmetry stating that private information deduced from trading causes market price fluctuations (See Kyle 1985) became important propositions. Consequently, many market attributes

have been employed as substitutes for information in the study of security price behaviour. Price changes occur in discrete increments, which are denoted in multiples of ticks. It is well recognised today that failing to treat the price process as a discrete series could adversely affect prediction results. Initially the modeling of discrete transaction prices was done by Gottlieb and Kalay (1985). The generalisation and variation of such a modeling framework can be found in Ball (1988); Glosten and Harris (1988); Harris (1990); Dravid (1991) and Hasbrouck (1999). Most often, earlier studies have treated price change as a continuous variable, primarily focusing on the unconditional distribution, ignoring the timing of transactions, which is irregular and random. The "ordered probit model", which was initially proposed by Aitchison and Silvey (1957) is a useful model for discrete dependent variables, which can take only a finite number of values with a natural ordering. Gurland et al. (1960) developed it further and later it was introduced into the social sciences by McKelvey and Zavoina (1975), which became an analytical tool in the financial market security price dynamics of micro-structure research. This could be used to quantify the effects of various factors on stock price movements, whilst accounting for discreteness in price changes and the irregular spacing of trades.

In an ordered probit analysis of the conditional distribution of price changes, Hausman et al. (1972) recognised the importance of accounting for discreteness, especially in intraday price movements. In such fine samples, the extent of price change is limited to a few distinct values, which may not be well approximated by a continuous state space. Their paper investigated the impact of several explanatory variables in capturing the transaction price changes. Importantly, the clock-time effect, measured in terms of duration between two consecutive trades, bid-ask spread, trade size and market-wide or systematic movements in prices based on a market index on conditional distribution of price changes were modeled under this framework. In a more recent study, Yang and Parwada (2012) extended the existing empirical literature on the impact of market attributes on price dynamics, utilising an ordered probit model. Their study explored the price impact of variables such as market depth and trade imbalance (also referred to as order imbalance in quote driven markets), in addition to trade size, trade indicator, bid-ask spread and duration which were found to be significant in similar studies. The model thus estimated by Yang and Parwada (2012), was able to forecast the direction of price change for about 72% of the cases, on average.

The in-sample and out-of-sample forecasts provided by the authors were based on the observed values of the regressors in the forecast horizon. However, in generating out-of-sample forecasts beyond one-step ahead incorporating observed values for regressors is of limited practical use, as they are not observed priori. Developing multi-step ahead forecasts, at least for a few transactions ahead is much more beneficial from a practical perspective, for effective decision making. However, such forecasting evidence under this framework is seemingly absent in the literature. Therefore, in addressing this shortcoming, this paper introduces a forecasting mechanism to generate forecasts beyond the one-step ahead level. Towards this end, disaggregated forecasts are generated first, for each of the explanatory variables for the period concerned. In order to generate forecasts for the regressors included, the serial dependence structure of each of the variables is investigated and appropriate forecasting models are fitted. Sign forecasts are subsequently generated, based on those predicted regressor values, rather than on observed values and the estimated coefficients of the ordered probit model. These prediction results are compared with those of the existing literature. Through the introduction of dynamic variables into the forecasting system, the predictive capability of this approach is investigated through a study based on the stocks of seven major companies listed in the Australian Securities Exchange (ASX).

In summary, the primary motivation of this paper is to introduce a method to enhance the flexibility and adaptability of the ordered probit model to generate multi-step ahead forecasts of stock price changes. Identifying and estimating appropriate univariate models for forecasting each explanatory variable, taking their serial dependence structure into account, towards this endeavour, is the second motivation. The third motivation is to improve on the results of Yang and Parwada (2012) in model estimation and forecast accuracy, by reducing noise in the data used and suitably formulating variables. Therefore, this exercise features the same stocks and almost the same independent variables that were

employed by Yang and Parwada (2012). We were able to achieve an 88 per cent plus rate of accuracy, on average, in the out of-sample forecasts of the direction of price changes using observed regressor values. In addition, more than 91 per cent of in-sample estimates, on average, correctly predicted the direction of price change. This is in comparison to the 72 per cent achieved by Yang and Parwada (2012). It is between 78-80 per cent when predictied regressor values were incorporated.

The remainder of the paper is organized as follows. Section 2 provides a review of the ordered probit model while Section 3 gives a description of the data and the variables used in the analysis. This section reports the summary statistics for each variable for the chosen stocks and introduces the relevant models for estimation and forecasting of durations, residuals and regressors. The empirical evidence is reported in Section 4 including model estimation and diagnostics. The results of the forecasting exercise for both in-sample and out-of-sample are presented in Section 5 and finally, the concluding remarks are provided in Section 6.

2. A Review of the Ordered Probit Model

In a sequence of transaction prices, P_{t_0}, P_{t_1}, P_{t_2}, …, P_{t_T} occurring at times t_0, t_1, t_2, …, t_T the resulting price changes multiplied by 100 is represented as an integer multiple of a tick and denoted by $Y_1, Y_2, \ldots, Y_T$, where $Y_k \equiv \{P_{t_k} - P_{t_{k-1}}\} \times 100$. The ordered probit model analyses discrete dependent variables with responses that are ordinal but not continuous. Underlying the indexing in such models, there exists a latent continuous metric and the thresholds partition the real line into a series of different regions corresponding to these ordinal categories. Therefore, the unobserved latent continuous variable Y^* is related to the observed discrete variable Y. It is assumed that the conditional mean of Y^* is described as a linear combination of observed explanatory variables, X and a disturbance term that has a Normal distribution.

The ordered probit specification takes the following form:

$$Y_k^* = X_k'\beta + \varepsilon_k, \quad \text{where} \quad \varepsilon_k | X_k \sim \ i.n.i.d.N(0, \sigma_k^2), \tag{1}$$

where i.n.i.d denotes that the errors are independently but not identically distributed. X_k is a $q \times 1$ vector of predetermined explanatory variables that govern the conditional mean, Y_k^* and β is a $q \times 1$ vector of parameters to be estimated. Here, the subscript denotes the transaction time. The observed price change Y_k is related to the latent continuous variable Y_k^* according to the following scheme:

$$Y_k = \begin{cases} s_1 & \text{if } Y_k^* \in A_1 \\ s_2 & \text{if } Y_k^* \in A_2 \\ \vdots & \vdots \\ s_m & \text{if } Y_k^* \in A_m \end{cases}, \tag{2}$$

where the sets A_k are comprised of non overlapping ranges of values, partitioning the continuous state space of Y_k^* and the s_j are the corresponding discrete values containing the state space of Y_k, which are called states. Let s_j's be the price change in ticks $-2, -1, 0, 1, \ldots$. Suppose that the threshold values of A are given as follows:

$$\begin{cases} A_1 \equiv (-\infty, \alpha_1], \\ A_2 \equiv (\alpha_1, \alpha_2], \\ \vdots \\ A_k \equiv (\alpha_{k-1}, \alpha_k], \\ \vdots \\ A_m \equiv (\alpha_{m-1}, \infty). \end{cases} \tag{3}$$

The number of states, m is kept finite, though in reality price change could take any value in cents to avoid the explosion of an unknown number of parameters. As per Hausman et al. (1972), the only requirement in this framework is the conditional independence of the ε_k's, where all the serial dependence would be captured by the regressors. Further, there are no restrictions on the temporal dependence of the X_k's. The conditional distribution of Y_k , conditioned upon X_k depends on the partition boundaries and the distributional assumption of ε_k. The conditional distribution in the case of Gaussian ε_k is

$$P(Y_k = s_i|X_k) = P(X_k'\beta + \varepsilon_k \in A_i|X_k)$$

$$= \begin{cases} P(X_k'\beta + \varepsilon_k \leq \alpha_1|X_k) & \text{if } i = 1, \\ P(\alpha_{i-1} < X_k'\beta + \varepsilon_k \leq \alpha_i|X_k) & \text{if } 1 < i < m, \\ P(\alpha_{m-1} < X_k'\beta + \varepsilon_k|X_k) & \text{if } i = m, \end{cases} \tag{4}$$

$$= \begin{cases} \Phi\left(\frac{\alpha_1 - X_k'\beta}{\sigma_k}\right) & \text{if } i = 1, \\ \Phi\left(\frac{\alpha_i - X_k'\beta}{\sigma_k}\right) - \Phi\left(\frac{\alpha_{i-1} - X_k'\beta}{\sigma_k}\right) & \text{if } 1 < i < m, \\ 1 - \Phi\left(\frac{\alpha_{m-1} - X_k'\beta}{\sigma_k}\right) & \text{if } i = m, \end{cases} \tag{5}$$

where $\Phi(\cdot)$ denotes the standard Normal cumulative distribution function. Since the distance between the conditional mean $X_k'\beta$ and the partition boundaries determines the probability of any observed price change, the probabilities of attaining each state, given the conditional mean, could be changed by shifting the partition boundaries appropriately. The explanatory variables capture the marginal effects of various economic factors that influence the likelihood of a given state as opposed to another. Therefore, the ordered probit model determines the empirical relation between the unobservable continuous state space and the observed discrete state space as a function of the explanatory variables, X_k, by estimating all the system parameters, including β coefficients, the conditional variance σ_k^2 and the partition boundaries α, from the data itself.

Let U_{ik} be an indicator variable, which takes the value 1 if the realisation of the kth observation, Y_k is the ith state s_i and 0 otherwise. The log likelihood function L for the price changes $Y = [Y_1, Y_2, \ldots, Y_T]$, conditional on the regressors, $X = [X_1, X_2, \ldots, X_T]$, takes the following form:

$$\begin{aligned} L(Y|X) = \textstyle\sum_{k=1}^{T} \Big\{ & U_{1k}.log\Phi\left(\frac{\alpha_1 - X_k'\beta}{\sigma_k}\right) \\ + \textstyle\sum_{i=2}^{m-1} U_{ik}.log & \left[\Phi\left(\frac{\alpha_i - X_k'\beta}{\sigma_k}\right) - \Phi\left(\frac{\alpha_{i-1} - X_k'\beta}{\sigma_k}\right)\right] \\ + U_{mk}.log & \left[1 - \Phi\left(\frac{\alpha_{m-1} - X_k'\beta}{\sigma_k}\right)\right] \Big\} \end{aligned} \tag{6}$$

Hausman et al. (1972) has reparameterised the conditional variance σ_k^2 based on the time between trades and lagged spread.

Models for Correlated Errors and Explanatory Variables

As mentioned in the above subsection, models with an appropriate autoregressive structure are used as forecasting models for the explanatory variables. Autoregressive integrated moving average (ARIMA) models of order (p,d,q) or ARIMA (p,d,q) models are used tomodel the autocorrelation in a time series and are used to predict behaviour based on past values alone. However, certain variables warranted the application of a simple ARIMA type model while others exhibit long range dependence,

which require autoregressive fractionally integrated moving average (ARFIMA) (p,d,q) type models to describe their behaviour. On the other hand, forecasts of indicator variables with more than two categories are based on multinomial logistic regressions, where the responses are nominal categories. The heteroscedasticity in the residuals is captured by the generalised autoregressive conditional heteroscedasticity GARCH(p, q) model (Bollerslev 1986), following (Yang and Parwada 2012). A brief description of each of these models are given in the Appendix.

3. Data, Variables and ACD Model

3.1. Data Description and ACD Model

The relevant data for this analysis was obtained from the Securities Industry Research Centre of Asia-Pacific (SIRCA) in Australia. The dataset consists of time stamped tick-by-tick trades, to the nearest millisecond and other information pertaining to trades and quotes for the chosen stocks listed in the Australian Securities Exchange (ASX). This study is based on a sample of stock prices collected during a three month period from 16 January 2014 to 15 April 2014. The stocks that were not subjected to any significant structural change, representing seven major industry sectors, are included in the sample. The selected stocks are Australian Gas Light Company (AGL), BHP Billiton (BHP), Commonwealth Bank (CBA), News Corporation (NCP), Telstra (TLS), Westfarmers (WES) and Woodside Petroleum (WPL) from Utilities, Materials, Financials, Consumer Discretionary, Telecommunication services, Consumer Staples and Energy sectors respectively. All seven of these stocks are included in the study by Yang and Parwada (2012), consisting of both liquid and less liquid assets, to minimise sample selection biases. However, the sampling period and the sample size differ between studies. Two stocks are not included in this paper due to the absence of transactions during the study period. Intraday price changes extracted from tick by tick trade data forms the basic time series under consideration. Overnight price changes are excluded as their properties differ significantly from those of intraday price changes (See Amihud and Mendelson 1987; Stoll and Whaley 1990). The trading hours of ASX are from 10.00 a.m. to 4.00 p.m. Due to the possibility of contamination of the trading process by including opening and closing trades (Engle and Russell 1998), the trades during the initial 30 min of opening and the final 30 min prior to closing are disregarded.

The following information with respect to each transaction is collected for each stock: Trade data comprising of date, time, transaction price and trade size, quote data such as bid price and ask price, market depth data comprising of volume at the highest bid price (best bid) and volume at the lowest ask price (best ask) and market index (ASX200). HFD generally contains erroneous transactions and outliers that do not correspond to plausible market activity. This is mainly attributed to the high velocity of transactions (Falkenberry 2002). Among others Hansen and Lunde (2006); Brownlees and Gallo (2006) and Barndorff-Nielsen et al. (2009) have paid special attention to the importance of data cleaning. A rigorous cleaning procedure is used here in obtaining a reliable data series for the analysis, mainly in accordance with the procedure outlined in Barndorff-Nielsen et al. (2009). To generate a time series at unique time points, during the instances of simultaneous multiple trades (quotes), the median transaction price (bid/ask prices) of those trades (quotes) is considered. Correspondingly, cumulative volume of those trades (quotes) are taken as the trade volume (bid/ask volume).

In the ordered probit model, the dependent variable Y_k is the price change between the kth and $k-1$th trade multiplied by 100. This records Y_k in cents, which however is equivalent to ticks as the tick size of the ASX for stocks with prices of the chosen magnitude is 1 cent. In this analysis, several different explanatory variables are included to measure their association with direction of price movement, following Yang and Parwada (2012). Bid and ask quotes are reported as and when quotes are updated, which necessitates the matching of quotes to transaction prices. Each transaction price is matched to the quote reported immediately prior to that transaction. Similarly, aggregate volumes at the best bid and best ask prices together with the ASX200 index representing the market are also matched in a similar fashion. The bid-ask spread $Sprd_{k-1}$, is given in cents, while $LBAV_{k-1}$ & $LBBV_{k-1}$ denote the natural

log of number of shares at best ask and bid prices respectively. $LVol_{k-1}$ gives the natural logarithm of $(k-1)$th trade size. Conditional duration, ψ_{k-1} and standardised transaction duration ϵ_{k-1} are derived estimates by fitting an autoregressive conditional duration model (ACD (1,1)) to diurnally adjusted duration data. A brief description of the model introduced by Engle and Russell (1998) is presented in Appendix A. The initial record of each day is disregarded as it is linked to the previous day's prices and results in negative durations. TI_{k-1} denotes the trade indicator of $(k-1)$th trade, which classifies a trade as buyer-initiated, seller-initiated or other type of trade. Trade imbalance TIB_{k-1}, based on the preceding 30 trades that occurred on the same day (Yang and Parwada 2012) (YP hereafter) is calculated as follows:

$$TIB_{k-1} = \frac{\sum_{j=1}^{30} \left(TI_{(k-1)-j} \times Vol_{(k-1)-j} \right)}{\sum_{j=1}^{30} Vol_{(k-1)-j}} \tag{7}$$

The first 30 observations of trade imbalance (TIB) is set to zero as TIB also depends on the previous day's trade imbalance for these transactions.

Market index return $RIndx_{k-1}$, prevailing immediately prior to transaction k is computed as given below:

$$RIndx_{k-1} = \ln(INDX_{k-1}) - \ln(INDX_{k-2}) \tag{8}$$

The sampling period and the use and categorisation of certain variables in this analysis differ from YP. ASX200 is applied here instead of specific sector indexes as the impact of the performance of the overall economy tends to be more significant on stock price behaviour than of a specific sector. On the other hand, the reference point for grouping the price changes is the 'one tick' threshold vis a vis the 'zero' change. This provides a more meaningful classification of the groups, as the categorisation of price change is based on a range of values rather than a fixed value for a certain group.

3.2. Sample Statistics

The main characteristics of the chosen variables in the analysis and how those characteristics differ between stocks could be ascertained from the several summary statistics that are provided in Table 1. There is considerable variation in the price level among the stocks considered in the sample. The highest price during this period ranged between AUD 4.96 for TLS and AUD 77.87 for CBA. The volatility of prices as indicated by the standard deviation of the percentage price change is not very high, with the TLS recording the highest value of 7.65 per cent. For most other stocks, it is less than 5 per cent. Average trade volume also records a substantial dispersion between the stocks, which varied from 161 for NCP and 6983 for TLS during the period. An indication of whether a transaction is buyer-initiated or seller-initiated is required for the empirical analysis. This measure is useful in identifying the party most anxious to execute the trade and the actions of whom would be reflected in terms of the bid/ask spread. The trades fall into these two categories in more or less equal proportions across stocks and are very similar in value except for TLS. The indeterminate trades form around 8–18% of trades, while it is 45% for TLS. The absence of asymmetric pressure from the buying or selling side suggests that there were no events with major news impact that would have resulted in abnormal trades and returns. This is further highlighted by zero mean returns.

The trading frequency as measured by the average duration between two consecutive trades also varies across stocks significantly. For more liquid stocks such as BHP, CBA and WES, trades tend to occur every 5 s or less on average. The other stocks are generally traded within 10 s. However, NCP is traded every 25 s on average. The observed large dispersions is a characteristic inherent in trade durations. Next, the estimation of the duration dynamics under an ACD model is considered, since the expected and standardised durations enter the orderd probit model as two separate variables.

The estimated coefficients of the ACD (1,1) model fitted to diurnally adjusted durations is presented in Table 2. The multiplicative error component is assumed to follow a Standardised Weibull distribution. All the coefficients are highly significant for each of the stocks, indicating the dependence

of the expected duration on its past behaviour. It is straightforward to estimate the conditional expected durations, ψ_k utilising the parameter estimates from the ACD model. The diurnal component was estimated using a cubic spline with knots at each half hour between 10:30 a.m. and 15:30 p.m. The standardised durations or the unexpected durations, ϵ_k are then obtained by dividing the diurnally adjusted durations by the conditional expected durations, which is an i.i.d. process. The parameter estimates are based on the conditional maximum likelihood approach, using the standardised Weibull distribution for ϵ_k. The Weibull distribution is a better choice here as opposed to exponential since the shape parameter is statistically significant and different from unity for all the stocks. Refer to the Appendix for the corresponding log-likelihood function.

Table 1. Descriptive statistics of the variables considered in the ordered probit model for all the stocks, for the period from 16 January 2014 to 15 April 2014.

Statistic	**AGL**	**BHP**	**CBA**	**NCP**	**TLS**	**WES**	**WPL**
Price (AUD)							
Max price	16.15	39.79	77.87	20.17	5.29	43.93	39.5
Min price	14.71	35.06	72.15	16.92	4.96	40.88	36.54
Price Change (%)							
Mean	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Std.dev	0.0376	0.0197	0.0124	0.0728	0.0765	0.0181	0.0183
Duration (Seconds)							
Mean	9.59	3.51	3.49	24.76	8.04	4.26	5.30
Std.dev	19.01	7.37	7.56	49.91	12.47	9.08	11.08
Trade Volume							
Mean	395	710	285	161	6983	281	318
Std.dev	2206	3622	2183	711	39,379	1370	1290
Shares at the Best Bid Price							
Mean	4451	5498	1579	877	941,002	1841	1983
Std.dev	5027	6464	2899	1994	603,015	2504	2406
Shares at the Best Ask Price							
Mean	4399	5513	1808	977	992,906	1945	2100
Std.dev	5409	7544	5053	1649	642,204	2775	2719
Market Index Returns, *ASX200*							
Mean	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Std.dev	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Trade Imbalance							
Mean	−0.0268	−0.0119	0.0094	−0.0623	0.0094	0.0224	0.0004
Std.dev	0.4653	0.4590	0.4401	0.4863	0.5082	0.4564	0.4446
Trade Direction (%)							
Buyer initiated	40.9	41.0	44.7	43.9	27.0	44.2	44.6
Seller initiated	41.6	41.0	42.2	48.1	27.6	40.6	41.9

The standardised durations are deemed weakly exogenous in the case of Australian stocks, according to the regression results of YP. They have regressed the standardised residuals on trades, volumes and returns for each of the stocks, which included the seven stocks of our study. On the other hand, both these studies consider the lagged measures of duration, addressing the problem of endogeneity to some extent. Furthermore, Dufour and Engle (2000) have treated durations as a strongly exogeneous variable in assessing the role of time on price dynamics.

The volumes at the best bid and ask prices prevailing prior to a transaction gives a measure of market depth. TLS has the deepest market, minimising the price impact cost for its trades. The trade imbalance (TIB) attempts to capture the cumulative demand side and supply side discrepancy over the last 30 trades. TIB < 0, if seller-initiated cumulative trading volume exceeded the buyer-initiated cumulative trading volume, during the immediately preceeding 30 trades prior to the current transaction. On the other hand, TIB > 0, if the buyer-initiated volume was more than the seller-initiated volume. The zero indicates either all indeterminate trades or an exact matching of selling and buying volumes during the period. In any case, zeros are very rare. Overall, there is a insignificant trade imbalance across all stocks. However, three stocks have a negative sign implying the selling volume marginally exceeded the buying volume while the other four stocks have a positive sign indicating the reverse phenomenon.

Table 2. The coefficient estimates of an ACD (1,1) model with Standardised Weibull errors fitted for the stocks. The conditional expected duration where x_k is the adjusted duration. α is the shape parameter of the Weibull distribution.

Parameter	AGL	BHP	CBA	NCP	TLS	WES	WPL
α_0	0.3177 (21.92 *)	0.0024 (10.16 *)	0.3178 (32.47 *)	0.0348 (10.03 *)	0.0291 (28.82 *)	0.3349 (23.33 *)	0.0030 (6.65 *)
α_1	0.3113 (27.91 *)	0.0110 (20.61 *)	0.2195 (41.03 *)	0.1476 (15.53 *)	0.2180 (59.84 *)	0.1652 (29.71 *)	0.0201 (18.48 *)
β	0.4764 (26.80 *)	0.9865 (1371.66 *)	0.4949 (40.08 *)	0.8524 (89.69 *)	0.7820 (214.64 *)	0.5220 (30.26 *)	0.9785 (747.16 *)
α	0.2523 (427.88 *)	0.4295 (726.73 *)	0.4258 (731.49 *)	0.4369 (255.94 *)	0.5756 (476.06 *)	0.4194 (672.39 *)	0.4046 (582.99 *)

* Significant at 99% level.

It is noticed that most of the variables exhibit serial correlation, with variables such as $LVol$, $LBBV$, $LBAV$, $Sprd$, TI and TIB showing strong serial dependence, for all stocks. For illustration, Figures 1–3 present the time series behaviour together with the acf and pacf for a few selected variables for a random stock, AGL.

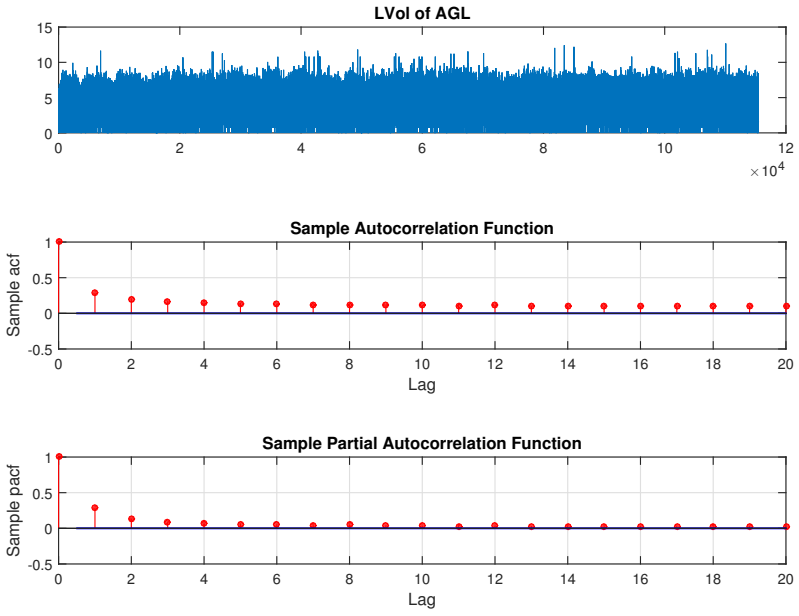

Figure 1. Time series, acf and pacf for LVol of AGL.

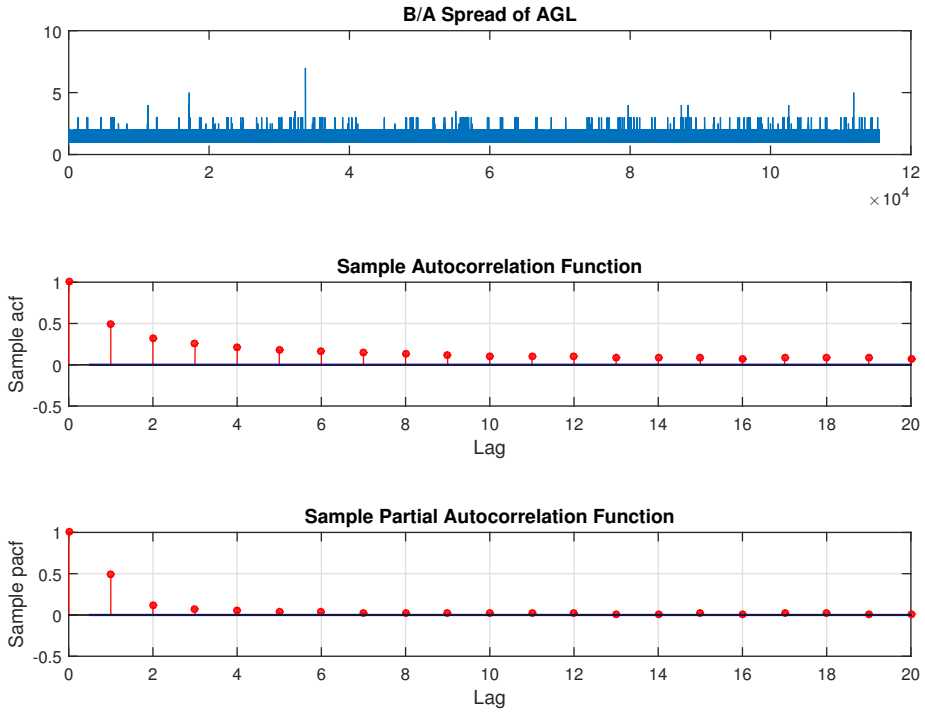

Figure 2. Time series, acf and pacf for spread of AGL.

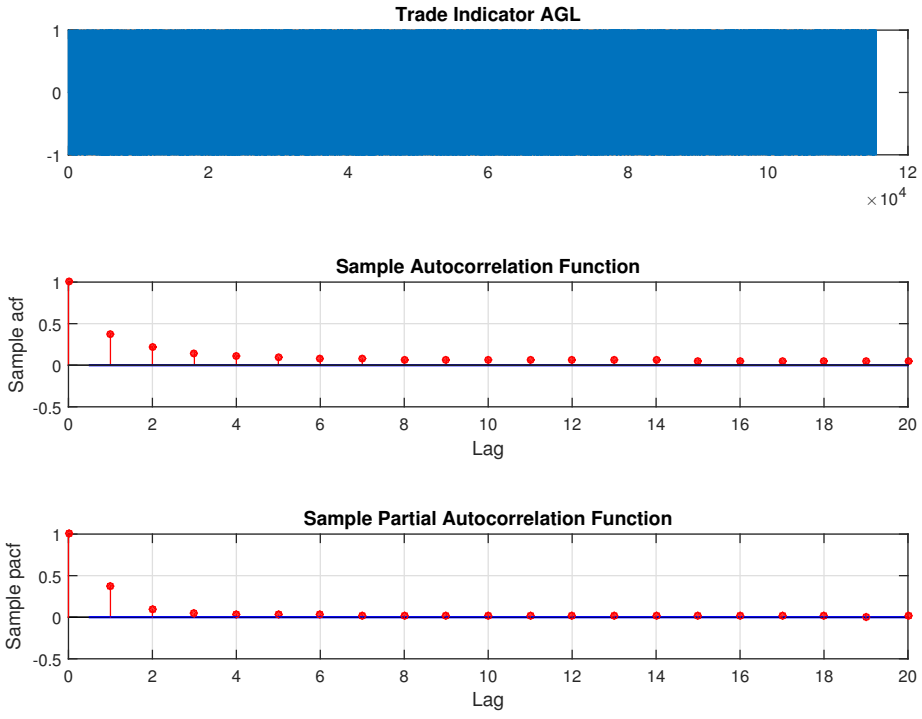

Figure 3. Time series, acf and pacf for trade indicator of AGL.

A novel feature of this study is that unlike in YP's study, we incorporate this feature by developing forecasting models for each explanatory variable based on the serial dependence structure. Therefore, Section 2 reports some useful models for capturing this feature in X_k's and σ_k^2, while details of the forecasting exercise is discussed later in Section 5.1.

4. Empirical Evidence

The model estimation for the direction of price change is carried out for these stocks for the period 16 January to 14 April 2014. Out of sample forecasts are generated for the last day of the sample, on 15 April 2014 from 10:30 a.m. to 15:30 p.m.

Y_k denotes the price changes between the k and $(k-1)$th trades in terms of integer multiples of ticks. The price change here is representative of the change in the observed transaction prices. The number of states that could be assumed by the observed price changes Y_k is set to 3, under the ordered probit framework. Price increases of at least 1 tick being grouped as +1, price decreases of at least 1 tick as -1, while price changes falling in $(-1,1)$, taking the value 0. The choice of m is based on achieving the balance between price resolution and minimising states with zero or very few observations. The decision to restrict m to 3 was mainly influenced by the fact that the observed price changes exceeding ± 2 ticks was below 0.05% for most stocks. The distribution of observed price changes in terms of ticks, over the transactions, is presented in Figure 4. Prices tend to remain stable in more than 80 per cent of the transactions, in general. For the rest of the time, rises and falls are more or less equally likely.

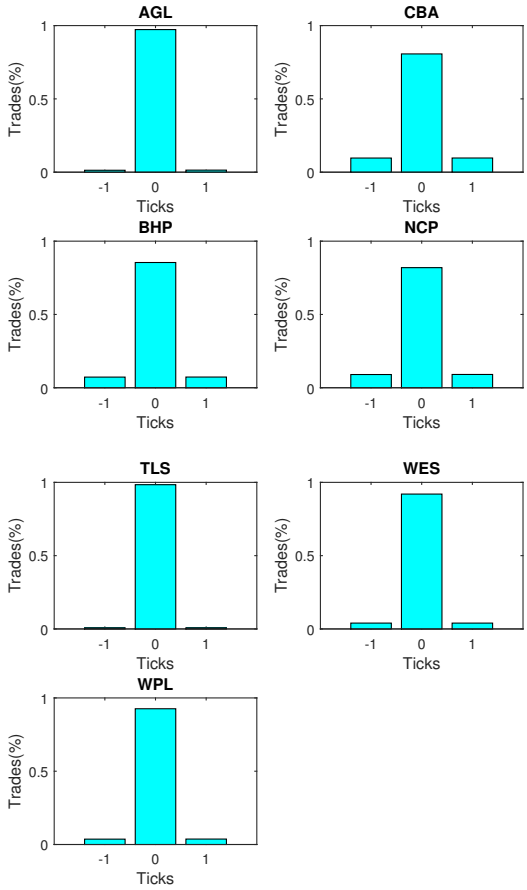

Figure 4. Distribution of the number of trades over the three categories of price change in terms of ticks, for all stocks during the period.

4.1. Ordered Probit Model Estimation

Prior to model estimation, all variables considered in the analysis was tested for stationarity using an Augmented Dickey-Fuller (ADF) test, which confirmed the same, which is in agreement with previous findings. The Ordered probit model specification depends on the underlying distribution of the price series. The model can assume any suitable arbitrary multinomial distribution, by shifting the

partition boundaries accordingly. However, the assumption of Gaussianity here has no major impact in deriving the state probabilities, though it is relatively easier to capture conditional heteroscedasticity.

The dependent variable in Equation (11) below is the price change in ticks. (An explanation of the latent continuous version of the price change was given in Section 2). The variables used in Equation (11) were described in Section 3. Just to recap, the first three variables on the R.H.S. of Equation (11) are three lags of the dependent variable. TI_{k-1} is the trade indicator; which classifies a trade as a buyer-initiated, seller-initiated or other type of trade. $SPRD_{k-1}$, is the Bid-Ask spread, measured in cents. $LVol_{k-1}$ gives the natural logarithm of $(k-1)$th trade size. $LBAV_{k-1}$ and $LBABV$ denote the natural log of number of shares at best ask and bid prices respectively. TIB_{k-1} is the trade imbalance, based on the preceding 30 trades on the same day. Conditional duration, ψ_{k-1} and standardised transaction duration ϵ_{k-1} are derived estimates by fitting an autoregressive conditional duration model (ACD (1,1)) to diurnally adjusted duration data. The ACD (1,1) model is described in the appendix. $RIndx_{k-1}$, prevailing immediately prior to transaction k, is calculated as the continuously compounded return on the ASX 200.

The mean equation under the ordered probit specification takes the following form:

$$\begin{aligned} X_k'\beta = \beta_1 Y_{k-1} + \beta_2 Y_{k-2} + \beta_3 Y_{k-3} + \beta_4 TI_{k-1} + \beta_5 Sprd_{k-1} + \beta_6 LVol_{k-1} \\ + \beta_7 LBAV_{k-1} + \beta_8 LBBV_{k-1} + \beta_9 TIB_{k-1} + \beta_{10} TI_{k-1} * \psi_{k-1} \\ + \beta_{11} TI_{k-1} * \epsilon_{k-1} + \beta_{12} RIndx_{k-1} \end{aligned} \tag{9}$$

The maximum likelihood estimates of the ordered probit model on price changes were computed based on BHHH algorithm of Berndt et al. (1974). The estimated coefficients of the above ordered probit system are presented in Table 3 while the corresponding z statistics are recorded within parentheses. Most of the regressors are highly significant to the model for all seven stocks, based on the asymptotically normally distributed z statistic (Hausman et al. 1972). The pseudo-R^2 values given at the bottom of the table show an improvement, irrespective of the number of observations, in comparison to those of YP. A relatively higher number of significant coefficients across all stocks is another improvement.

The first three lags of the dependent variable comes under scrutiny, first. All the lags are significant with a 95% confidence level, with a negative coefficient for each stock. This inverse relationship with past price changes is consistent with the existing literature, indicating a reversal in the price compared to its past changes. Consider a one tick rise in price over the last three trades in the case of AGL, for example, keeping the other variables constant. The subsequent fall in the conditional mean (Y_k^*) would be 3.9448, which is less than the lower threshold, resulting in -1 for Y_k. The coefficients of the traditional variables such as the bid ask spread ($Sprd$), trade volume (LVol) and the market index returns are significant for all stocks but one, in each case. The $Sprd$ and $LVol$ has a positive impact on the price change across all stocks. The market index returns, based on the ASX200, as a measure of the overall economy, generally has a significant positive impact on price changes. Overall, this is in line with the conventional wisdom. Meanwhile, the coefficients of the trade indicator, the number of shares at the best bid price and the number of shares at the best ask price are significant for all stocks.

The trade imbalance (TIB) between buyers and sellers has a positive impact on price change and is statistically significant across all stocks. This phenomenon agrees well with the general inference that more buyer-initiated trades tend to exert pressure from the demand-side, resulting in a subsequent rise in price and vice versa. The impact of the time duration between trades is measured separately via the two constituent components of an ACD model. One is the conditional expected duration (signed), $TI_{k-1} * \psi_{k-1}$ and the other isthe standardised innovations (signed), also referred to as unexpected durations, $TI_{k-1} * \epsilon_{k-1}$. The signed conditional expected duration is significant for all stocks while the unexpected component is significant for all but one. This highlights the informational impact of time between trades in price formation. The interpretation of these measures of duration is not straightforward as they are comprised of two components. The kind of impact those variables have on price change will depend on the significance of the trade initiation as well as on the durations.

One striking feature is that either both the components have a positive impact or both have a negative impact for a given stock. Wald tests were performed to investigate the significance of duration on price changes. The tests were conducted under the null hypotheses in which either the coefficient of the conditional duration is zero or the coefficient of standardised duration is zero or both are jointly zero. The resultant F statistics suggest that both the components of duration are significant for all the stocks considered. The test results are not presented here for the sake of brevity.

The partition boundaries produced below the coefficient estimates determine the partition points of the direction of change in the latent variable. There are three possible directions the price change can take in terms of ticks, $Y_k \leq -1$, $Y_k = 0$ and $Y_k \geq +1$. By comparing these boundary values with the estimated continuous variable Y_k^*, values -1, 0 or +1 are assigned to the observed variable $\hat{Y}_k$.

Table 3. Coefficient estimates β_i, of ordered probit model on direction of price change based on 12 explanatory variables for the selected stocks. The sampling period was 16 January 2014 to 15 January 2014. Z statistics are given within parentheses for each parameter.

Parameter	AGL	BHP	CBA	NCP	TLS	WES	WPL
Obs.	*114,318*	*316,547*	*317,761*	*41,085*	*137,323*	*260,954*	*205,651*
Y_{k-1}	−2.2281 (−72.21 *)	−1.0240 (−130.16 *)	−0.7915 (−119.64 *)	−0.7637 (−54.32 *)	−1.7745 (−53.16 *)	−1.4750 (−143.59 *)	−1.6261 (−113.69 *)
Y_{k-2}	−1.1262 (−37.38 *)	−0.3614 (−50.67 *)	−0.3247 (−52.35 *)	−0.2554 (−18.92 *)	−0.9070 (−21.15 *)	−0.7413 (−72.42 *)	−0.8578 (−62.50 *)
Y_{k-3}	−0.5905 (−19.79 *)	−0.1142 (−16.65 *)	−0.1153 (−19.56 *)	−0.1405 (−10.50 *)	−0.4252 (−12.21 *)	−0.3533 (−33.69 *)	−0.4137 (−30.20 *)
TI	0.9572 (67.69 *)	1.2730 (285.32*)	−0.2832 (−88.07 *)	−0.1256 (−8.95 *)	−1.2319 (−38.61 *)	0.8899 (165.00 *)	0.9831 (146.16 *)
$Sprd$	0.0479 (3.16 *)	0.0342 (6.71 *)	0.0078 (2.60 *)	0.0117 (1.85 **)	0.0703 (1.21)	0.0238 (0.95)	0.0261 (4.83 *)
$LVol$	0.0047 (1.16)	0.0090 (6.13 *)	0.0050 (3.67 *)	0.0167 (4.92 *)	0.0118 (3.04 *)	0.0060 (3.03 *)	0.0043 (1.85 **)
$LBAV$	0.0801 (12.36 *)	0.1019 (48.39 *)	0.0337 (20.31 *)	−0.0312 (−6.94 *)	−0.0515 (−4.76 *)	0.0398 (15.25 *)	0.0607 (19.05 *)
$LBBV$	−0.0760 (−12.03 *)	−0.1097 (−53.04 *)	−0.0397 (−23.81 *)	0.0252 (5.66 *)	0.0425 (3.32 *)	−0.0575 (−21.15 *)	−0.0744 (−21.51 *)
TIB	0.0488 (2.52 *)	0.0647 (9.70 *)	0.0929 (15.89 *)	0.0457 (3.04 *)	0.1916 (9.47 *)	0.0986 (11.50 *)	0.0696 (6.77 *)
$TI * \psi$	−0.2246 (−31.79 *)	−0.3764 (−119.95 *)	0.7304 (203.38 *)	0.0353 (4.24 *)	−0.0546 (−3.91 *)	−0.2506 (−62.24 *)	−0.2631 (−56.41 *)
$TI * \epsilon$	−0.0535 (−11.51 *)	−0.0608 (−37.49 *)	0.0706 (55.74 *)	0.0013 (0.40)	−0.0213 (−4.01 *)	−0.0314 (−14.58 *)	−0.0328 (−12.73 *)
$RIndx$	262.0988 (3.30 *)	110.1514 (2.47 *)	194.2312 (3.71 *)	307.72 (7.49 *)	−107.5324 (−0.87)	342.8466 (6.52 *)	139.1455 (2.72 *)
α_1	−2.8628	−2.1469	−1.6552	−1.5723	−4.5651	−2.3399	−2.4639
α_2	2.9999	2.1869	1.6699	1.7676	5.1375	2.2403	2.3768
$Pseudo - R^2$	0.3203	0.3339	0.2226	0.2068	0.3223	0.2589	0.2833

* Significant at 95% level. ** Significant at 90% level.

In parameterising the conditional variance, an ARMA specification was used following YP. Therefore, a GARCH (p, q) specification including up to two lags was used on the residual series of the ordered probit model across all stocks. The orders p, q were selected on the basis of Akaike information criterion (AIC). The selected parameter estimates of the fitted GARCH models are reported in Table 4. Only some of the parameters appear to be significant with less persistence in conditional volatility for some stocks.

Table 4. Coefficient estimates of GARCH parameters of the conditional variance of the residuals for all stocks. ω, constant; κ, GARCH parameters; δ, ARCH parameters

Parameter	AGL	BHP	CBA	NCP	TLS	WES	WPL
Obs.	*114,318*	*316,547*	*317,761*	*41,085*	*137,323*	*260,954*	*205,651*
ω	0.4081 (0.3419)	0.0468 (0.5782)	0.0022 (0.4748)	0.0598 (0.3415)	0.0204 (0.1226)	0.3242 (0.5995)	0.3282 (0.4998)
κ_1	0.3194 (0.2182)	0.3803 (0.3205)	0.9723 (46.8638)	0.8255 (2.4207)	0.9175 (2.9561)	0.2445 (0.2668)	0.2406 (0.2209)
κ_2		0.5157 (0.4516)				0.2045 (0.2272)	0.2172 (0.2019)
δ_1	0.2523 (0.6870)	0.0429 (0.9682)	0.0244 (1.4103)	0.0873 (0.6766)	0.0615 (0.0.2994)	0.1735 (1.3347)	0.1660 (1.1188)

4.2. Price Impact of a Trade

Price impact measures the effect of a current trade of a given volume on the conditional distribution of the subsequent price movement. In order to derive this, $X_k\beta$ has to be conditioned on trade size and other relevant explanatory variables. The volumes, durations and the spread were kept at their median values while the index was fixed at 0.001 whereas trade indicator and trade imbalance were kept at zero to minimise any bias. It is observed that the coefficients of the three lags of Y_k are not identical, implying path dependence of the conditional distribution of price changes (Hausman et al. 1972). Consequently, the conditioning has to be based on a particular sequence of price changes as well, as a change in the order will affect the final result. These conditioning values of X_k's specify the market conditions under which the price impact is to be evaluated.

The conditional probabilities were estimated under five scenarios of path dependence keeping the other quantities at the specified values. These are falling prices $(-1/-1/-1)$, rising prices $(1/1/1)$, constant prices $(0/0/0)$ and alternative price changes, $(-1/+1/-1)$ and $(+1/-1/+1)$. Figures 5 and 6 exhibit the plots of estimated probabilities under the first three scenarios for all the seven stocks. The shifts in the distribution are clearly evident for the first two cases as against the third case of constant prices. Under the falling price scenario, the shift is more towards the right while for the rising price scenario, it is more towards the left indicating an increased chance of price reversal after three consecutive rises or falls. In the case of alternating prices it was revealed that prices tend to remain stable in the subsequent trade.

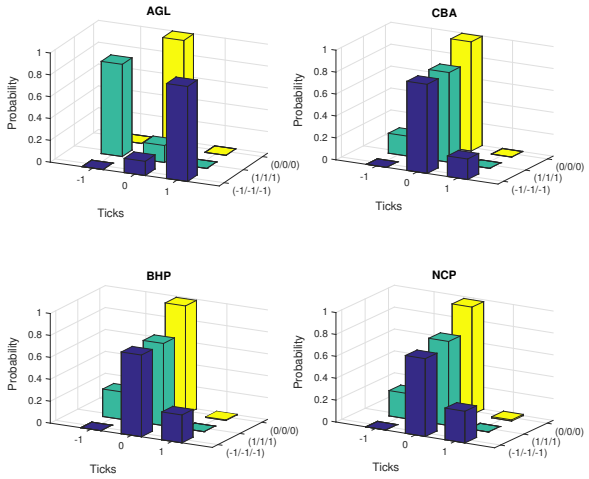

Figure 5. Distribution of estimated probabilities of direction of price change conditioned on constant, increasing and decreasing past price changes.

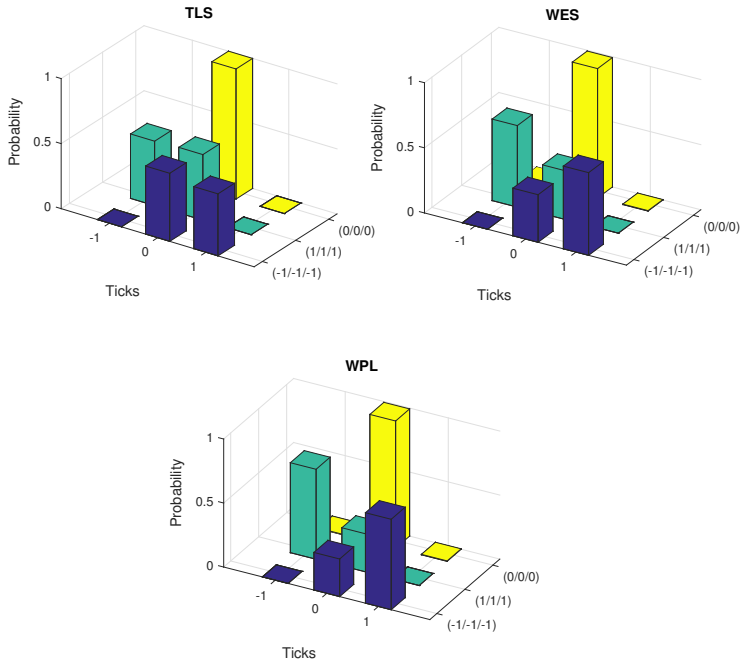

Figure 6. Distribution of estimated probabilities of direction of price change conditioned on constant, increasing and decreasing past price changes.

4.3. Diagnostics

A well specified ordinary least squares (OLS) regression would exhibit little serial correlation in the residuals. A similar kind of test could be performed on the generalised residuals in the case of ordered probit to test its validity, as it is not possible to obtain residuals directly (Hausman et al. 1972). Table 5 contains the sample cross-correlation coefficients of generalised residuals with the lagged generalised fitted values, $\hat{Y}_{k-j}$, computed up to 12 lags. Under the null hypothesis of no serial correlation, the theoretical cross-correlation coefficients should be zero or close to zero. The reported values are quite small, varying in the range from -0.01 to 6.19×10^{-6}.

Table 5. Cross-autocorrelation coefficents $\hat{v}_j, j = 1, \ldots, 12$ of generalised residuals with lagged generalised fitted price changes.

Parameter	AGL	BHP	CBA	NCP	TLS	WES	WPL
$\hat{v}_1$	-0.0025	-0.0002	-0.0015	-0.0004	0.0004	-0.0015	-0.0004
$\hat{v}_2$	-0.0057	0.0012	-0.0015	-0.0002	-0.0012	-0.0003	0.0009
$\hat{v}_3$	-0.0103	-0.0005	-0.0016	0.0008	-0.0008	0.0010	0.0015
$\hat{v}_4$	-0.0058	6.19×10^{-6}	-0.0018	-0.0029	-0.0028	-0.0004	0.0013
$\hat{v}_5$	-0.0045	0.0006	-0.0018	-0.0022	-0.0039	0.0005	-0.0017
$\hat{v}_6$	-0.0056	-0.0001	-0.0020	-0.0025	0.0016	0.0031	0.0018
$\hat{v}_7$	0.0009	-0.0008	-0.0018	0.0002	-0.0015	0.0034	0.0001
$\hat{v}_8$	0.0029	0.0001	-0.0017	0.0043	-0.0039	0.0010	0.0003
$\hat{v}_9$	0.0001	-7.76×10^{-5}	-0.0017	0.0057	-0.0023	0.0036	-0.0023
$\hat{v}_{10}$	0.0047	0.0003	-0.0015	0.0039	-0.0021	0.0030	0.0042
$\hat{v}_{11}$	0.0076	0.0020	-0.0013	0.0025	-0.0033	0.0009	0.0017
$\hat{v}_{12}$	0.0011	0.0014	-0.0011	0.0014	-0.0041	0.0024	-0.0002

5. Forecasting the Direction of Price Change

The forecasting performance of the ordered probit model fitted to the stocks is investigated. The tests of in-sample and out of-sample forecasts provide some basis to gauge the model's ability to accurately forecast the future direction of price changes. Forecasts are generated under three scenarios. In-sample probability estimates are based on the last week of the training sample from 8 April to 14 April 2014. Meanwhile, out-of-sample forecasts are based on the final day of the data series, 15 April.

Only one day is considered for the out-of-sample performance as it is not feasible to project price changes beyond one day with any degree of accuracy as a normal trading day contains more than 1000 transactions for all the stocks, with the exception of NCP which had only 417. Out-of-sample forecasts are computed in two ways. One is one-step ahead forecasts based on the observed, recorded values of the regressors and the other is the multi-step ahead, using their predicted values. The next subsection discusses the forecast generation under the second scenario in more detail. The commonly observed measures of forecast performance are not so relevant in this case, since the dependent variable is categorical. However, some measures such as root mean square error (RMSE) and mean absolute deviation (MAD) were calculated for both in-sample and out-of-sample forecasts, though they are not reported here for the sake of brevity.

5.1. Out-of-Sample Multi-Step Ahead Forecasts with Disaggregated Predictions of Individual Explanatory Variables

In real life, the values of the regressors are not observed priori, to forecast at least a few transactions ahead. Unlike in YP's study this paper develops out-of-sample multi-step ahead forecasts based on disaggregated predictions of the regressors. Under this scenario, multi-step ahead forecasts are generated for the entire forecast horizon, based on the estimated models, as well as 100-step ahead rolling basis. The rolling forecasts of price change are based on similar forecasts of explanatory variables. Towards this end, we first predict the future values of the regressors based on models that are fitted to capture the autoregressive behaviour of each variable in the sample. Under this setup, forecasts of price change are derived for the estimated transactions occurring on the last day of the series, 15 April 2014. The relevant models are fitted after a careful inspection of the autocorrelation function (acf) and the partial autocorrelation function (pacf) of the individual series, as discussed in Section 2. The model selection among several competing models is based on the AIC for a given regressor. In most instances, the time series of *LVol* shows a hyperbolic decay in their acfs and pacfs, similar to Figure 1. Therefore, an ARFIMA type model is the preferred choice for *LVol*. The fractional differencing parameter, d is always within the range of 0 to 0.5, indicating the presence of long memory. On the other hand, most other variables such as $LBBV$, $LBAV$, TIB and $Sprd$ have slow decaying autocorrelations and partial autocorrelations, with the majority falling short of a hyperbola. Figure 2 gives a general perception on the behaviour observed in these variables. For these regressors, an ARMA type model suffices for most stocks, in general. Forecasts of trade indicator are based on a multinomial logistic regression on $LBAV$, $LBBV$, lags of Y and lags of TI, as the common contenders for the explanatory variables. Parameter estimates of predictive models for selected variables are illustrated in Tables 6 and 7 for the stock, AGL. The expected and unexpected durations are forecasted by the estimated ACD model.

Table 6. Coefficient estimates of autoregressive model parameters fitted to selected independent variables. The t statistics are given within parentheses. Illustrative examples include a long memory and a short memory model for *LVol* and *Sprd* for the stock AGL. d, long memory parameter; ϕ, AR parameters; θ, MA parameters.

Parameter	LVol	Spread
	(ARFIMA)	**(ARMA)**
c		0.0030 (7.30)
d	0.1867 (68.50)	
ϕ_1	0.0082 (15.72)	1.7555 (160.732)
ϕ_2		−0.7581 (−71.06)
θ_1	−0.0079 (−35.59)	−1.3455 (−123.55)
θ_2		0.2774 (41.16)
θ_3		0.0621 (15.35)
θ_4		0.0136 (3.67)
θ_5		0.0052 (1.80)

Table 7. Coefficient estimates of multinomial logistic regression model parameters fitted to Trade indicator (TI) of AGL. The base category is 1. Z statistics are given in parentheses.

Independent Variable	Category	
	−1	0
c	0.1449 (2.63)	−1.9155 (−26.87)
dp_{k-1}	0.2095 (5.08)	0.1128 (1.98)
$lbbv_{k-1}$	−0.3146 (−60.57)	−0.0746 (−10.41)
$lbav_{k-1}$	0.3001 (60.42)	0.2302 (35.96)
TI_{k-1}	−0.9664 (−119.26)	−0.4632 (−42.96)

5.2. Forecast Performance of the Ordered Probit Model

The basic test of forecast errors is mainly based on the number of correct forecasts as a percentage of total forecasts. The fitted directions of price change, $\hat{Y}_k$, based on the estimated coefficients are compared with their actual counterparts for each transaction in the forecasting sample. The number of exact tallies provide the number of correct forecasts. The in-sample forecast results illustrated in Table 8 reports a 91% accuracy, which is a very high percentage, by any means, vouching for the significant forecasting ability of the model. In comparison, YP achieved a percentage of 72. On the other hand, out-of sample results are provided in Table 9. For one-step ahead forecasts based on observed regressor values, the direction could be accurately predicted 88 per cent of the time, on average, across all stocks. The percentage achieved by YP again is 72 per cent. Meanwhile, the performance of the multi-step ahead forecasts based on the fitted regressor values is not as striking as in the other two cases, as expected. Notwithstanding, percentages of 78 and 85, on average, are highly noteworthy and are still higher than the 72 per cent of YP. The comparatively dismal performance of TLS under the first scenario given in panel 2 (a) of Table 9 may have been influenced by a relatively small number of price changes recorded during the period. However, the rolling forecasts show a remarkable improvement. The ex-post forecast of this stock is slightly better than the ex-ante forecast, which is quite contrary to the other stocks. The reverse is observed for five of the other stocks, as anticipated, while for one stock, it is similar.

The predictions of regressors based on serial correlation structures do not provide very good long term multi-step ahead forecasts, due to mean reversion. As a result of this, the forecasts of price change direction, based on those fitted values may also not provide reliable long term forecasts. A single day is referred to as longterm as the average daily transactions exceed 1000 for most stocks in the sample. Therefore, under these circumstances, the forecast horizon is restricted to the 100 transactions of the last day on a rolling basis, which resulted in a much better accuracy percentage of 85, in comparison to the one incorporating all the transactions of that day. It is worthwhile mentioning that from an individual stock's perspective, the short term performance is better than the long term. The worst case scenario gives around 75 per cent of out-of-sample correct forecasts, whereas it is around 85 per cent for the in-sample predictions.

Based on predicted price movements, investors can adjust their trading positions accordingly in formulating trading strategies, risk management, portfolio allocations etc. However, the most risky position under these forecast scenarios would be the adverse selection (see Yang and Parwada 2012, for more details). It is where the actual occurrence is the opposite of the predicted price movement, with possible adverse effects on the investor's networth. Therefore, it is worthwhile examining the extent of the possibility of this risk of adverse selection taking place. The percentages of predictions in the opposite direction for actual rise/fall are given in Tables 8 and 9 for in-sample predictions and out-of-sample forecasts respectively. Generally, this risk is very small and not more than 1 per cent across all stocks, except TLS, under all the forecast scenarios. In the case of out-of-sample forecasts, TLS records a 50 per cent risk of adverse selection, mainly as a result of only two recorded price falls in the forecast sample. Furthermore, altogether there are only three rises/falls in the price, giving rise to zero correct classifications for those categories for TLS.

The predicted conditional probabilities of the three categories of forecast price change $\hat{Y}_k$, -1, 0 and 1 are generated under the ordered probit system for in-sample as well as out-of-sample forecasts. $\hat{Y}_k$ is assigned the value of the category with the highest probability for a given transaction. These probabilities obtained for the stock, CBA, are illustrated for 100 observations during the forecast period, in Figures 7 and 8 to represent all the stocks, which show similar behaviour. For a given observation, the vertical sum of the three conditional probabilities is one. In the case of both in-sample and out-of-sample scenarios, the probabilities tend to fluctuate. However, for the majority of observations, no price change category tends to have a probability greater than 50 per cent, in general, resulting in lower percentages of correct classifications for rises and falls in prices. This does not indicate a deviation from the real life behavior in prices, with respect to the overall distribution across the three categories. Nevertheless, this phenomenon highlights a slight over prediction in that category. A similar pattern of behaviour is observed for the forecasts with predicted regressors as well.

As discussed earlier, most of the trades do not witness heavy movements in prices. Nevertheless, if a rise/fall in price could be foreseen in advance, investors are in a better position to create profitable strategies or to manage risk appropriately. Since multi-step ahead predictions of opposite price movements are rare, a forecast rise/fall would provide useful signals of future price directions. This, when combined with the knowledge of past price paths, will aid the investor in making a more informed decision in strategy formulation in his favour, especially towards minimising risk. However, improving the individual forecasts of the explanatory variables will be beneficial in realising better predictions of future price movements under this framework.

Table 8. In-sample predictions of direction of price change for the last one week period of the training sample from 8 April 2014 to 14 April 2014.

Parameter		**AGL**		**BHP**		**CBA**		**NCP**		**TLS**		**WES**		**WPL**	
One Week—08/04–14/04															
Observations		114,318		316,547		317,761		41,085		137,323		260,954		205,651	
accuracy(%)		97.80		86.20		85.11		85.24		98.75		92.19		94.76	
-1		40.00		23.79		36.03		25.48		39.29		27.11		38.45	
0		99.79		98.67		98.14		98.64		99.65		99.51		99.78	
+1		45.40		24.29		35.73		32.01		36.00		33.57		37.09	
Actual	Forecast	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
-1	+1	0	0	1	0.05	6	0.32	1	0.42	0	0	0	0	0	0
+1	-1	0	0	0	0	3	0.17	0	0	0	0	0	0	0	0

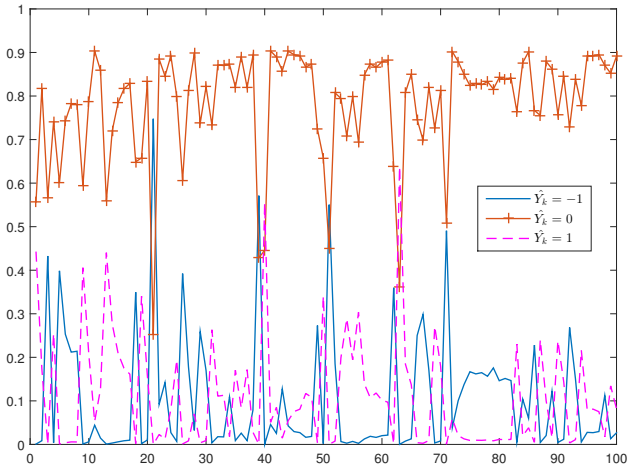

Figure 7. In-sample estimated probabilities of direction of price change for 100 observations of CBA.

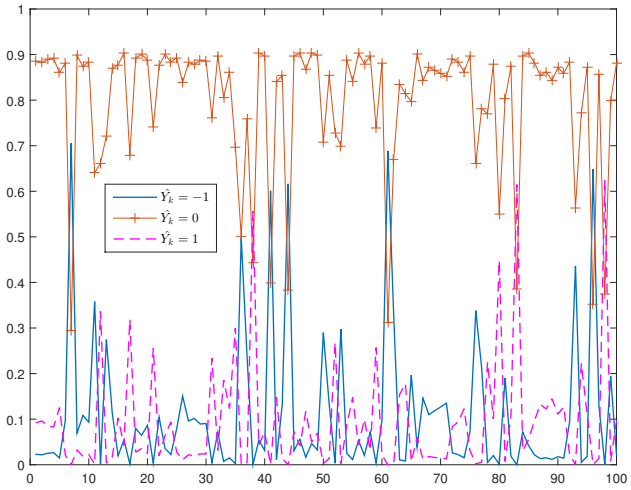

Figure 8. Out of-sample estimated probabilities of direction of price change for 100 observations of CBA, based on actual regressor values.

Table 9. Out-of sample forecasts of direction of price change for the last day of the sample, 15 April 2014. First panel contains one-step ahead forecasts based on actual explanatory variables and the second panel, muti-step ahead with predicted variables.

Parameter		**AGL**	**BHP**	**CBA**	**NCP**	**TLS**	**WES**	**WPL**
One-Step Ahead—15/04								
Observations		1151	3319	4073	417	1218	2518	3638
accuracy (%)		97.83	85.90	79.65	75.30	99.67	88.28	91.70
−1		35.29	27.40	26.80	21.54	0.00	32.54	32.77
0		99.73	98.47	97.31	96.82	99.92	98.95	99.47
+1		38.89	23.84	26.88	37.68	0.00	38.03	35.53
Actual	Forecast	No. %	No. %	No. %	No. %	No. %	No. %	No. %
−1	+1	0 0	0 0	6 1.2	0 0	0 0	2 1.2	0 0
+1	−1	0 0	0 0	1 0.2	0 0	0 0	0 0	0 0
Multi-Step Ahead—15/04								
(a) All transactions								
Observations		1151	3319	4073	417	1218	2518	3638
accuracy (%)		97.74	82.74	74.93	67.87	47.46	87.41	90.07
(b) 100-step ahead								
accuracy (%)		97.48	83.25	80.55	74.82	79.97	89.87	92.25
−1		30.77	25.34	21.94	20.00	0.00	23.78	20.43
0		98.93	94.94	99.87	100.00	80.16	99.14	99.88
+1		38.46	29.18	23.72	23.19	0.00	27.45	25.71
Actual	Forecast	No. %	No. %	No. %	No. %	No. %	No. %	No. %
−1	+1	0 0	2 0.68	3 0.58	0 0	1 50	0 0	0 0
+1	−1	0 0	0 0	0 0	0 0	0 0	0 0	0 0

6. Conclusions

The future direction of stock price movements are predicted through the estimation of an ordered probit model under an empirical setup. The study comprises of intra-day transaction data of seven stocks representing seven industry sectors, listed on the ASX. The ordered probit specification seems to adequately capture the price changes. All the explanatory variables are highly significant for the majority of the stocks. Diagnostics indicate lack of serial correlation in residuals with the implication of the model providing a good fit. The sequence of trades has an impact on the conditional distribution of price changes, while the trade size too is important with larger volumes putting more pressure on prices.

In improving forecast accuracy of the model, our study differs from that of YP in certain respects. Percentage of success in predicting future direction of price movements is used as a yardstick to measure the forecasting strength of the model. The success rate of the in-sample predictions is around 91 per cent and out-of sample one-step ahead forecasts happens to be 88 per cent. These percentages are much higher than the respective percentage of 72 per cent achieved by YP, for both cases. Overall, our forecasts outperform those of YP for each common stock.

Another main contribution of this study is to forecast price changes within a more practical perspective. In real life, the values of the regressors are not known a-priori to forecast at least a few transactions ahead. In addressing this drawback, we first predict the future values of the regressors based on their serial correlation structures by way of appropriate models. This resulted in several Autoregressive Moving Average (ARMA) and Autoregressive Fractionally Integrated Moving Average (ARFIMA) type models. In a subsequent step, these disaggregated forecasts are incorporated into the ordered probit model to generate future price change forecasts. Obviously, the 100-steps ahead short term forecasts perform better than the longterm ones including all transactions in the forecast horizon for most stocks. On average, the successful percentage in the long term is still a reasonable 78 per cent, which is affected by a poorly performing stock. On the other hand, the average success rate in the short term is around 85 per cent, which is quite remarkable.

Given the considerably high percentage of constant prices in real life, the model captures this phenomenon, albeit with a slight bias towards predicting no change. However, the risk of adverse selection is minimised. Nevertheless, this predictive model is useful for investors in developing successful trading strategies, particularly towards minimising risk as this provides valuable signals towards the future directions of price movements. The usefulness of this model to growth driven investors could be enhanced by improving the forecasting accuracy of the independent variables by adopting more sophisticated econometric techniques within a unified framework. In addition, the investigation of the adequacy of the conditional variance specification may also prove useful in improving the forecast probabilities.

Author Contributions: R.Y., R.G., S.P. and D.E.A. conceived and designed the experiments. R.Y. performed the experiments and analysed the data. R.Y., R.G., S.P. and D.E.A. wrote the paper.

Acknowledgments: The authors thank the reviewers for helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Models for Errors and Explanatory Variables

Appendix A.1. ACD Specification to Model Transaction Durations

The Autoregressive Conditional Duration (ACD) model analyses transaction duration, identifying it as a conditional point process.The temporal dependence of the diurnally adjusted duration process is captured by the conditional expected duration, $\psi_k = E(x_k|x_{k-1}, \ldots, x_1)$, under a linear ACD$(p, q)$ specification and has the following form:

$$\psi_k = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{k-i} + \sum_{j=1}^{q} \beta_j \psi_{k-j}, \tag{A1}$$

where $p \geq 0; q \geq 0$. The standardized durations

$$\epsilon_k = \frac{x_k}{\psi_k}$$

are i.i.d. with $E(\epsilon_i) = 1$. The log likelihood function for the Std. Weibull errors, is

$$L(x|\theta) = \sum_{k=2}^{T} \alpha \ln\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right] + \ln\left(\frac{\alpha}{x_k}\right) + \alpha \ln\left(\frac{x_k}{\psi_k}\right) - \left[\Gamma\left(1+\frac{1}{\alpha}\right)\frac{x_k}{\psi_k}\right]^{\alpha}, \tag{A2}$$

where α is the shape parameter of the density.

Appendix A.2. GARCH Specification to Model Heteroscedasticity

Conditional variance in the residuals (ε_k) of the orderd probit model, σ_k^2 can be estimated from this model.

$$\varepsilon_k = \sigma_k \eta_k, \quad \varepsilon_k | X_k \sim (o, \sigma_k^2)$$

where σ_k^2 is the conditional volatility of ε_k given the past historical information, η_k is a sequence of independently and identically distributed (i.i.d.) random variables with zero mean and variance 1 such that

$$\sigma_k^2 = \omega + \sum_{i=1}^{p'} \kappa_i \sigma_{k-i}^2 + \sum_{j=1}^{q'} \delta_j \varepsilon_{k-j}^2.$$

Appendix A.3. ARIMA Model

A series $\{X_k\}$ that could be modeled as a stationary ARMA(p'', q'') process after being differenced d times is denoted as ARIMA(p'', d, q'') with the following form:

$$\phi(B)(1-B)^d X_k = c + \theta(B) a_k, \tag{A3}$$

where $\phi(B) = 1 - \phi_1 - \ldots - \phi_p$, $\theta(B) = 1 + \theta_1 + \ldots + \theta_q$ and $a_k \sim WN(0, \sigma_a^2)$ and B is the backshift operator.

Appendix A.4. Long Memory ARFIMA Model

ARFIMA is designed to capture the long range dependence in time series. This model extends the ARIMA model in (A3) allowing d to lie between -0.5 and + 0.5 yielding a fractionally integrated series. ARFIMA process is said to exhibit stationary long memory if $d \in (0, 0.5)$. See Granger and Joyeux (1980) for details.

An ARFIMA(p'', d, q'') process has the same form as in (A3) and the operator $(1-B)^d$ is given by

$$(1-B)^d = \sum_{k=0}^{\infty} \frac{\Gamma(k-d) B^k}{\Gamma(-d)\Gamma(k+1)}; \quad d \notin \{1, 2, \ldots\}$$

Appendix A.5. Multinomial Logistic Regression

Multinomial logistic regression is an extension of binary logistic regression, that handles polytomous responses. This is used to predict the response category or the probability of category membership of a nominal outcome variable. The log odds of the outcome are modeled as a linear combination of multiple explanatory variables.

If $x_k = (x_{k1}, x_{k2}, \ldots, x_{kr})'$ follows a multinomial distribution with r response categories and parameter $\pi_k = (\pi_{k1}, \pi_{k2}, \ldots, \pi_{kr})'$, then

$$log\left(\frac{\pi_{kj}}{\pi_{kj}^*}\right) = y_k^T \beta_j', j \neq j^*$$

considering j^* as the baseline category. Assuming that mth category is the baseline category ($j^* = m$), the coefficient vector is

$$\beta' = (\beta_1', \beta_2', ..., \beta_{m-1}', \beta_{m+1}', ..., \beta_r')$$

References

Aitchison, John, and Samuel D. Silvey. 1957. The generalization of probit analysis to the case of multiple responses. *Biometrika* 44: 131–40. [CrossRef]

Amihud, Yakov, and Haim Mendelson. 1987. Trading mechanisms and stock returns: An empirical investigation. *The Journal of Finance* 42: 533–53. [CrossRef]

Ball, Clifford A. 1988. Estimation bias induced by discrete security prices. *The Journal of Finance* 43: 841–65. [CrossRef]

Barndorff-Nielsen, Ole E., P. Reinhard Hansen, Asger Lunde, and Neil Shephard. 2009. Realized kernels in practice: Trades and quotes. *The Econometrics Journal* 12: C1–32. [CrossRef]

Berndt, Ernst R., Bronwyn H. Hall, Robert E. Hall, and Jerry A. Hausman. 1974. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* 3: 653–65.

Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]

Breen, William, Lawrence R. Glosten, and Ravi Jagannathan. 1989. Economic significance of predictable variations in stock index returns. *Journal of Finance* 44: 1177–89. [CrossRef]

Brownlees, Christian T., and Giampiero M. Gallo. 2006. Financial econometric analysis at ultra-high frequency: Data handling concerns. *Computational Statistics & Data Analysis* 51: 2232–45.

Cheung, Yin-Wong, Menzie D. Chinn, and Antonio Garcia Pascual. 2005. Empirical exchange rate models of the nineties: Are any fit to survive? *Journal of International Money and Finance* 24: 1150–75. [CrossRef]

Christoffersen, Peter F., and Francis X. Diebold. 2006. Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science* 52: 1273–87. [CrossRef]

Dravid, Ajay R. 1991. *Effects of Bid-Ask Spreads and Price Discreteness on Stock Returns*. Technical Report. Philadelphia: Wharton School Rodney L. White Center for Financial Research.

Dufour, Alfonso, and Robert F. Engle. 2000. Time and the price impact of a trade. *Journal of Finance* 55: 2467–98. [CrossRef]

Engle, Robert F., and Jeffrey R. Russell. 1998. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* 66: 1127–62. [CrossRef]

Falkenberry, Thomas N. 2002. *High Frequency Data Filtering*. Great Falls: Tick Data Inc.

Glosten, Lawrence R., and Lawrence E. Harris. 1988. Estimating the components of the bid/ask spread. *Journal of Financial Economics* 21: 123–42. [CrossRef]

Gottlieb, Gary, and Avner Kalay. 1985. Implications of the discreteness of observed stock prices. *The Journal of Finance* 40: 135–53. [CrossRef]

Granger, Clive W. J., and Roselyne Joyeux. 1980. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1: 15–29. [CrossRef]

Gurland, John, Ilbok Lee, and Paul A. Dahm. 1960. Polychotomous quantal response in biological assay. *Biometrics* 16: 382–98. [CrossRef]

Hansen, Peter R., and Asger Lunde. 2006. Realized variance and market microstructure noise. *Journal of Business & Economic Statistics* 24: 127–61.

Harris, Lawrence. 1990. Estimation of stock price variances and serial covariances from discrete observations. *Journal of Financial and Quantitative Analysis* 25: 291–306. [CrossRef]

Hasbrouck, Joel. 1999. The dynamics of discrete bid and ask quotes. *The Journal of Finance* 54: 2109–42. [CrossRef]

Hausman, Jerry A., Andrew W. Lo, and A. Craig MacKinlay. 1992. An ordered probit analysis of transaction stock prices. *Journal of Financial Economics* 31: 319–79. [CrossRef]

Kyle, Albert S. 1985. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society* 53: 1315–35. [CrossRef]

Leung, Mark T., Hazem Daouk, and An-Sing Chen. 2000. Forecasting stock indices: A comparison of classification and level estimation models. *International Journal of Forecasting* 16: 173–90. [CrossRef]

Madhavan, Ananth. 2000. Market microstructure: A survey. *Journal of Financial Markets* 3: 205–58. [CrossRef]

McKelvey, Richard D., and William Zavoina. 1975. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* 4: 103–20. [CrossRef]

Pesaran, M. Hashem, and Allan Timmermann. 2004. How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting* 20: 411–25. [CrossRef]

Stoll, Hans R., and Robert E. Whaley. 1990. Stock market structure and volatility. *Review of Financial Studies* 3: 37–71. [CrossRef]

White, Halbert. 2000. A reality check for data snooping. *Econometrica* 68: 1097–126. [CrossRef]
Yang, Joey Wenling, and Jerry Parwada. 2012. Predicting stock price movements: An ordered probit analysis on the Australian Securities Exchange. *Quantitative Finance* 12: 791–804. [CrossRef]