Applications of Machine Learning to Estimating the Sizes and Market Impact of Hidden Orders in the BRICS Financial Markets

Maake, Witness and Van Zyl, Terence

28 February 2020

Applications of Machine Learning to Estimating the Sizes and Market Impact of Hidden Orders in the BRICS Financial Markets

Witness R. Maake[1], Terence L. Van Zyl[1,a]

[a] *Wits Institute of Data Science, Department of Computer Science and Applied Mathematics, University of the Witwatersrand, 1 Jan Smuts Avenue, Braamfontein, 2000, Johannesburg, South Africa*

Abstract

The research aims to investigate the role of hidden orders on the structure of the average market impact curves in the five BRICS financial markets. The concept of market impact is central to the implementation of cost-effective trading strategies during financial order executions. The literature of Lillo *et al.* (2003) is replicated using the data of visible orders from the five BRICS financial markets. We repeat the implementation of Lillo *et al.* (2003) to investigate the effect of hidden orders. We subsequently study the dynamics of hidden orders. The research applies machine learning to estimate the sizes of hidden orders. We revisit the methodology of Lillo *et al.* (2003) to compare the average market impact curves in which true hidden orders are added to visible orders to the average market impact curves in which hidden orders sizes are estimated via machine learning. The study discovers that : (1) hidden orders sizes could be uncovered via machine learning techniques such as Generalized Linear Models (GLM), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forests (RF); and (2) there exist no set of market features that are consistently predictive of the sizes of hidden orders across different stocks. Artificial Neural Networks produce large R^2 and small $\hat{\text{MSE}}$ on the prediction of hidden orders of individual stocks across the five studied markets. Random Forests produce the most appropriate average price impact curves of visible and estimated hidden orders that are closest to the average market impact curves of visible and true hidden orders. In some markets, hidden orders produce a convex power-law far-right tail in contrast to visible orders which produce a concave power-law far-right tail. Hidden orders may affect the average price impact curves for orders of size less than the average order size; meanwhile, hidden orders may not affect the structure of the average price impact curves in other markets. The research implies ANN and RF as the recommended tools to uncover hidden orders.

Keywords:
Hidden Orders;
Market Features;
GLM; ANN; SVM; RF;
Hidden Order Sizes;
Market Impact;
BRICS(Brazil, Russia, India, China, and South Africa).

JEL Classification : C45, C55, G12, G14, G15, G32

* Corresponding W. Maake. Tel.: +27 79 480 1825 . Email address: 363807@students.wits.ac.za (W. Maake)
** Corresponding T. Van Zyl. Tel.: +27 11 717 6073 . Email address: terence.vanzyl@wits.ac.za (T. Van Zyl)

1. Introduction

The discipline of Market Microstructure goes back to the discovery by Maureen O'Hara who defined it as the study of the process and outcomes of exchanging assets under explicit trading rules. The discipline deals with issues of market structure and design, price formation and price discovery, transaction and timing costs, information and disclosure, market maker and investor behavior. O'Hara *et al.* (1995)'s broad market micro-structure theoretical framework leads to numerous publications of concepts that are related to the operations of financial markets. Lillo *et al.* (2003) is one such publication.

Lillo *et al.* (2003) reveals the volume dependence of price impact of single trades on the New York Stock Exchange (NYSE), and reports a power-law dependence, $\Delta P(\lambda, \omega_i, \alpha) = \frac{1}{\lambda} sign(\omega_i)|\omega_i|^{\alpha}$, where λ is the market liquidity proxy derived from the mean market capitalization of the stocks that were analyzed, ω_i is the normalized trade size of trade i, $P(\lambda, \omega_i, \alpha)$ is the price change due to the trade of size ω_i, and α is the slope of the price impact curves that is a scalar in the range $[0.1, 0.4]$.

Fellow researchers such as Harvey *et al.* (2016) argued that perhaps if the Lillo *et al.* (2003) was to be implemented in a different market with small changes on the choice of parameters, then different results could be observed. Based on the research of Lillo *et al.* (2003) and Harvey *et al.* (2016), we argue diligently that if hidden orders were to be accounted for on the calculations of average market impact curves, then different structures of the average market impact curves could emerge.

We acknowledge the contributions of Lillo *et al.* (2003) and Harvey *et al.* (2016) for providing feasible research that allows our research to be formulated. We proceed to investigate the micro-structures of five BRICS financial markets. We reveal the average market impact of single trades in each of the BRICS financial markets. We establish a research genre that allows us to declare our research legitimate by replicating the research of the literature. We formulate three research questions : (1) What role does market capitalization play on the structure of the average market impact curves of the five BRICS financial markets?, (2) Do hidden orders affect the structure of the average market impact curves in the five BRICS markets?, and (3) Can machine learning assist us to uncover the hidden characteristics of hidden orders, in particular the sizes of such hidden orders?

We formulate research objectives that allow us to answer the research questions unambiguously: (1) We investigate the role of market capitalization on the structure of market impact, (2) We investigate the role of hidden orders on the structure of market impact, and (3) We train machine learning techniques to estimate hidden orders.

Our machine learning techniques are regression-based supervised learning algorithms. Inputs and output are known and are related through well defined models. The chosen models in the research context are Generalized Linear Models, Artificial Neural Networks, Support Vector Machines and Random Forests. The learning task entails the determination of the parameters of each of the models that would better explain the relationship between inputs and output.

2. Methodology

We studied roughly nine months of data for each market from the inception date of 28 - 09 - 2018 to the maturity date of 28 - 06 - 2019. The data for each market is comprised of historic trades and quotes that were sourced from the Thomson Reuters Eikon. The Johannesburg Stock Exchange (JSE) data set is comprised of 63 firms with a total of 10, 975, 007 transactions. The Bolsa de Valores de Sao Paulo (BOVESPA) data set includes 42 firms with a total of 6, 572, 183 transactions. The Moscow Exchange (MOEX) data set contains 30 firms with a total of 3, 091, 547 transactions. The National Stock Exchange of India (NSE) data set includes 43 firms with a total of 4, 095, 817 transactions. Finally the Shanghai Stock Exchange (SSE) data is comprised of 54 firms with a total of 5, 226, 211 transactions.

To attain our first objective we calculated the average daily value traded per firm, as per Harvey *et al.* (2016), which we applied as market capitalization proxy of each firm in each market. The periodic market capitalizations of firms that are reported at the closing dates of the financial periods of firms would be a loose approach to judging the performance of firms when we consider the possibility that firm ranks may change in real time. Henced we applied the average daily value traded as an appropriate parameter to classify firms into large capitalization and small capitalization in each market. We determined the median average daily value traded in each market. Firms with average daily value traded equal to or greater than the median are regarded as large-capitalization firms; otherwise, small-capitalization firms. We brief unfamiliar researchers on the average daily value traded. The product of the share price and the number of shares traded yields the value traded for the time instance. The average of such values traded over a trading day yields the average daily value traded for the day. The average of daily averages yields the average daily value traded over the trading days of consideration. We marked each firm as either large-capitalization or small-capitalization. We concluded our first objective by implementing the methodology of Lillo *et al.* (2003) over the data of visible orders.

In order to attain our second objective we identified hidden orders. We separated hidden orders from visible orders. We implemented the methodology of Lillo *et al.* (2003) on the data of visible orders. We re-implemented the methodology of Lillo *et al.* (2003) with aggregated data of visible and hidden orders to investigate the effect of hidden orders (Figures 1 - 5). Thereafter, we investigated the effect of hidden orders by studying the mean impact curves of seller and buyer initiated trading to within a 95% confidence interval from the mean (Figures A.21 - A.25).

The third objective was the contribution that distinguishes the research from the literature. We identified a set of features which we call market features (Appendix A). We mined the market features from the trading data using the appropriate methods of the literature. Using the market features we trained machine learning techniques namely : Generalized Linear Models (GLM), Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Random Forests (RF) to estimate the sizes of normalized volumes. We applied the trained machine learning techniques to estimate the sizes of hidden orders. We aggregated the estimated hidden orders with visible orders, and re-implemented the methodology of Lillo *et al.* (2003). We constructed the average market impact curve of visible and machine estimated hidden orders for each market to within a 90% confidence interval. We separately constructed the average market impact curve of visible and true hidden orders for each market also to within a 90% confidence interval. We fixed the scale at a specified domain and range suitable for each market and merged the results onto that scale to enable comparison.

We provide the details of the machine learning techniques to the unfamiliar researcher in Sections 2.1 to 2.4.

2.1. *Generalized Linear Models (GLM)*

We assumed that the noise factors, ϵ_i, were uncorrelated, independent and identically normally distributed with zero mean and constant variance, σ^2. We defined, in accordance with Seber *et al.* (1977), the generalized linear model as follows :

$$y_{i-1} = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_p x_{ip} + \epsilon_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m f_m(x_{i1}, x_{i2}, \ldots, x_{ip}) + \epsilon_i \tag{1}$$

where y_i is the time i response, α_m is the m^{th} coefficient, ϵ_i is the noise factor, x_{ij} is the time i j^{th} explanatory variable realization, $j = 1, \ldots, p$, $i = 1, \ldots, n$, $f(\cdot)$ is scalar-valued function that resembles a variety of forms such as polynomials and nonlinear functions. To maintain a credible reference frame of comparison, we deployed only 70% of the in-sample data to training. Of the 30% of sample data remaining, we deployed only 15% for testing. The remaining 15% was set aside for validation if required before testing. We present the results of the model in Table 1 - 10.

2.2. Artificial Neural Networks (ANN)

Following Nguyen *et al.* (1990), we fitted a two-layer feed-forward network with a sigmoid transfer function in the hidden layer, and a linear transfer function in the output layer. We subdivided the in-sample data into 70% training, 15% validation and 15% testing sets. The training set was deployed to determine a generalized input-output relationship. The validation set was used to impose the early-stopping of the training when generalization stopped improving. The testing set provided an independent measure of network performance. We set the initial number of neurons in the hidden layer to ten, which we may increase or decrease depending on the network performance per stock. We trained the network via the Levenberg - Marquardt backpropagation algorithm which is partly introduced in Marquardt *et al.* (1963). We present the results of the model in Table 11 - 15.

2.3. Support Vector Machines (SVM)

Conceptually, we applied the idea of Vapnik *et al.* (1995). Of the j market features namely : trade price, bid price, bid size, ask price, ask size, turnover ratio, price change, spread, mid quote price, average value traded, volatility, momentum, order sign, signed order, liquidity, order imbalance, divergence, execution time, and PIN, which are denoted by , x_{ij}, that are quantified at time i, we defined the response feature normalized volume, which we denoted by, ω_i. We determined the function $f_i(x_{i1}, x_{i2}, ..., x_{ip})$ such that the distance between ω_i and the function f_i is less than some well defined tolerance measure ϵ_i, and at the same time $f_i(x_{i1}, x_{i2}, ..., x_{ip})$ is as flat as possible. In closed form we minimize

$$\mathbf{J}(\alpha) = \frac{1}{2}\alpha'\alpha \tag{2}$$

subject to

$$\left|\omega_i - f_i(x_{i1}, x_{i2}, ..., x_{ip})\right| \le \epsilon_i \tag{3}$$

where

$$f_i(x_{i1}, x_{i2}, ..., x_{ip}) = \alpha_0 + \sum_{m=1}^{M} \alpha_m f_m(x_{i1}, x_{i2}, \dots, x_{ip}) = \alpha\mathbf{x} + \alpha_0. \tag{4}$$

However, we ought to state that a function that satisfied the two requirements for all evaluated points may not be satisfied. In which case we introduced the slack variable η_i and η_i^* for each point to allow regression error to bypass the ϵ_i threshold into $\epsilon_i + \eta_i$ threshold soft margin. With the additional conditions, our new objective is to minimize

$$\mathbf{J}(\alpha) = \frac{1}{2}\alpha'\alpha + \mathbf{A}\sum_{i=1}^{N}(\eta_i + \eta_i^*) \tag{5}$$

subject to

$$\forall i : \omega_i - f_i(x_{i1}, x_{i2}, ..., x_{ip}) \le \epsilon_i + \eta_i \tag{6}$$

$$\forall i : f_i(x_{i1}, x_{i2}, ..., x_{ip}) - \omega_i \le \epsilon_i + \eta_i^* \tag{7}$$

$$\forall i : \eta_i^* \ge 0 \tag{8}$$

$$\forall i : \eta_i \ge 0 \tag{9}$$

where $\mathbf{A} \in \mathbb{R}^+$ is the regularization parameter which determines the degree of trade-off between the amount by which deviations over the tolerance ϵ_i are accepted and the flatness of the function $f_i(x_{i1}, x_{i2}, ..., x_{ip})$. Considering the fact that the function $f_i(x_{i1}, x_{i2}, ..., x_{ip})$ is non linear due to its adaption to the non linear series of normalized volume, we drawn on the idea of Vapnik *et al.* (1995) to introduce Lagrange dual formula. Meanwhile, we replaced the linear dot product $x'_{i1}x_{i2}$ with the kernel function $\mathbf{G}(x_{i1}, x_{i2}) = \langle\psi(x_{i1}), \psi(x_{i2})\rangle$ where ψ is a mapping that transforms to high dimensional space. We determined the Gram matrix **G** directly and the optimal solution in the predictor space. The idea which is supported by Huang *et al.* (2005). In closed form we minimize

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_i - \alpha_i^*)\mathbf{G}(x_i, x_j) + \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{N}\omega_i(\alpha_i - \alpha_i^*) \tag{10}$$

subject to

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \tag{11}$$

$$\forall i : 0 \leq \alpha_i < \mathbf{A} \tag{12}$$

$$\forall i : 0 \leq \alpha_i^* < \mathbf{A} \tag{13}$$

with the subsequent Karush Kuhn Tucker conditions:

$$\forall i : 0 \leq \alpha_i(\epsilon + \eta_i - \omega_i + f_i(x_{i1}, x_{i2}, ..., x_{ip})) = 0 \tag{14}$$

$$\forall i : 0 \leq \alpha_i^*(\epsilon + \eta_i^* + \omega_i - f_i(x_{i1}, x_{i2}, ..., x_{ip})) = 0 \tag{15}$$

$$\forall i : \eta_i(\mathbf{A} - \alpha_\mathbf{i}) = \mathbf{0} \tag{16}$$

$$\forall i : \eta_i^*(\mathbf{A} - \alpha_\mathbf{i}^*) = \mathbf{0}. \tag{17}$$

All observations inside the epsilon radius posses the multipliers of zero, and the observations outside posses non-zero multipliers. They are regarded as support vectors which are solely the dependent variables of the function which is applied to predict new values which are provided by :

$$f_i(x_{i1}, x_{i2}, ..., x_{ip}) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\mathbf{G}(x_i x_{ip}). \tag{18}$$

Of the twenty-two features, we applied as inputs we refine our training approach, following Fan *et al.* (2005), to determine the minimal number of features that would obtain optimal training outcomes. Akin to the training of neural networks and generalized linear regression techniques we deployed 70% of the in-sample data to train the support vector machine model and 15% to test the model performance. We present the results of the model in Table 16 - 20.

2.4. *Random Forests (RF)*

In line with the literature of Breiman *et al.* (2001) we grew a regression bagged ensemble of about 200 trees. We sampled all variables at each node. To counteract the loss of accuracy due to missing information we specified surrogate splits. The split predictors were determined via the interaction test. The R^2 of the model was determined via out-of-bag predictions. We determined the significance of each market feature by permuting out-of-bag observations; otherwise, we could have determined the significance of each market feature by summing the gains in the mean squared error due to splits on each market feature. In line with the data splits of the generalized linear model, neural network and support vector machines, we split the in-sample data once again as 70% training and 15% of the remaining 30% for testing the performance of the random forest model. We present the results of the model in Table 21 - 25.

2.5. *Training and Validation*

The normalized volume was retrieved as the response variable; meanwhile, the other market features were injected as input variables. Of the total transactions that each market possessed, we split the data into halves. One half, referred to as the in-sample data, was applied for training and preliminary testing to deduce the terminal satisfactory parameter values. Each trained machine learning technique provided the implied statistics of good fit namely the R^2 and the

$\hat{\text{MSE}}$. The remaining half, the out of sample data, was used to verify the R^2 and $\hat{\text{MSE}}$ from the testing perspective. The $\hat{\text{MSE}}$ was calculated as the mean of squares of differences between the estimated normalized volumes and the targeted normalized volumes. The numerical values of the results indicated similarity to the R^2 and the $\hat{\text{MSE}}$ from the training perspective. We discovered empirically that machine learning techniques tend to exploit a unique set of features for each stock when attempting to estimate the sizes of hidden orders (Figures 6 - 10).

2.5.1. *Generalized Linear Models. Copyright 2013-2014 The MathWorks, Inc.*

The generalized linear models yielded: (1) the coefficient of a unit contribution of each feature in a linear model, (2) the standard error (SE), (3) the t-statistics (t-Stat), (4) the p-value associated with each feature, (5) the ordinary R^2, (6) the adjusted R^2, (7) the mean squared error ($\hat{\text{MSE}}$); and (8) the root mean squared error ($\text{R}\hat{\text{MSE}}$). The p-value is the parameter that allows us to judge the significance of each feature relative to others per firm; and across different firms. The performance of generalized linear models (GLM) appeared to be fairly acceptable as supported by the R^2 and the $\hat{\text{MSE}}$ parameters which are provided in Table 1 - 10. A closer analysis suggests that there exists no set of market features that are consistently predictive of the sizes of hidden orders across different stocks (Figures 6 - 10).

2.5.2. *Artificial Neural Networks. Copyright 1992-2015 The MathWorks, Inc.*

The artificial neural networks yielded: (1) the best epoch, (2) the best network performance, (3) the best validation performance, (4) the best test performance, (5) the R^2; and (6) the model mean squared error. The performance of artificial neural networks appeared to be exceptional. Evidence is provided by the extreme high R^2 and low $\hat{\text{MSE}}$ in Table 11 - 15. Other notable occurrences that transpired over the training were: (1) an increase in-network layers that captured the nonlinear relationship among the order sizes; however, that occurred at a trade-off of the linear relation, and (2) increasing or decreasing neurons in the hidden layer lead to good performance on the training and poor performance on the testing. Consequently, we made use of 2 layers and 10 neurons.

2.5.3. *Support Vector Machines. Copyright 2015-2016 The MathWorks, Inc.*

The support vector machines yielded: (1) the scalar feasibility gap between the dual and primal objective functions, (2) the scalar gradient difference between upper and lower violators, (3) the maximal scalar Karush-Kuhn-Tucker (KKT) violation value, (4) the numeric value of the dual objective, (5) the bias term in the SVM regression model; and (6) half the width of the epsilon-insensitive band. We standardized the data to allow the features to have a normalized frame of contribution such that comparison could be deduced. Numerous observations regarding the performance of Support Vector Machines suggested that SVM tend to exploit the nonlinear relationship among hidden orders, leading to the best performance over the GLM and the ANN in some instances. Empirical evidence is provided in Table 16 - 20.

2.5.4. *Random Forests. Copyright 2013-2016 The MathWorks, Inc.*

The random forests yielded: (1) the re-substitution loss, (2) the R^2; and (3) the $\hat{\text{MSE}}$. Akin to SVM the random forests exploited the non linearity relationship in hidden orders well, which also has to lead to its best performance over the GLM and the ANN occasionally (Table 21 - 25). We also attempted to train the random forest ensemble via the LS boosting of Freund *et al.* (1997) and the G boosting of Friedman *et al.* (2001). The bagged regression ensemble produced better performance mainly because of its relevance to the structure of the data and its optimization routines. We chose the bagged regression as the appropriate way to estimate hidden orders.

3. Results and Discussion

3.1. *South Africa (JSE)*

We observe that small average daily value traded firms tend to cluster at the top; meanwhile, the large average daily value traded firms tend to cluster at the bottom (Figure 1). Similar observations were reported in Lillo *et al.* (2003). We comment confidently on the research question of the role of market capitalization on the structure of

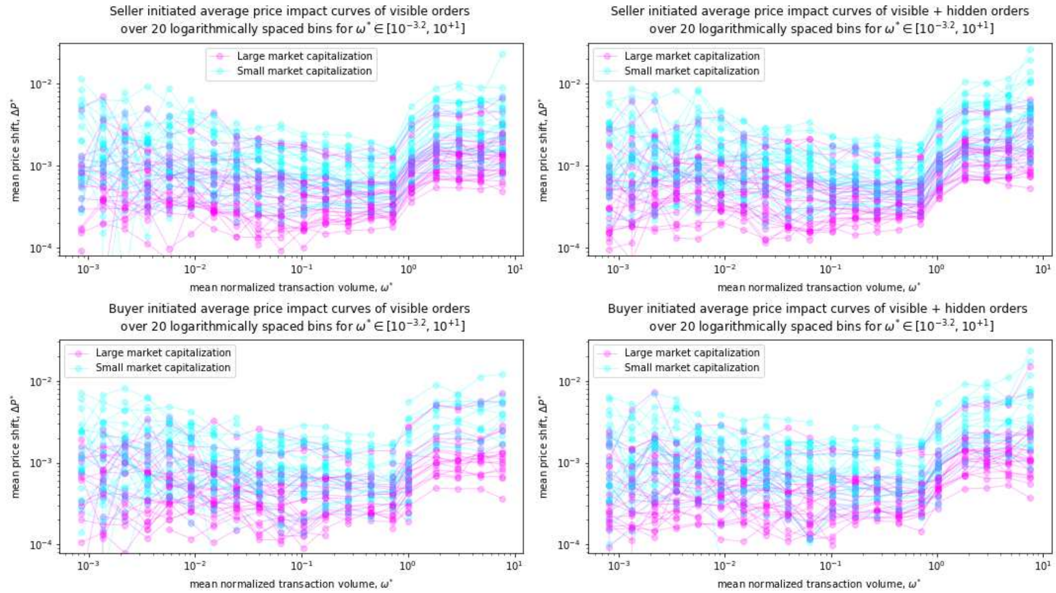

Fig. 1. Price impact curves of single trades of 63 liquid firms. The curves are computed using the analytical literature approach. (Magenta Curves) represent firms with the average daily value traded larger than the median average daily value traded. (Cyan Curves) represent firms with the average daily value traded smaller than the median average daily value traded.The analysis is period from 28-09-2018 to 28-06-2019.

Johannesburg Stock Exchange (JSE). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

the average market impact curves. Market capitalization does not change the structure of the average market impact curves; however, it distinguishes firms that are costly to trade and firms that are less costly to trade. Other notable observations of Figure 1 are that the average price shift lies roughly in between the range $[10^{-4.1}, 10^{-1.5}]$. Harvey *et al.* (2016) reported similar price shift bounds. Figure A.21 provides the results of buyer initiated and seller initiated trading of visible orders (left) and visible & hidden orders (right). We held the standard deviation constant to allow our analysis to answer the research question of the impact of hidden orders on the average market impact curves. However, the advanced procedure would be to allow the standard deviation to vary to enable possible comments on how confident we are with our interpretation of the results. We respond to how hidden orders affect the impact curves in a casual sense rather than to delve deeper into assessing the degree to which our interpretation is true. Figure A.21 shows that hidden orders effect a strict convex power-law far-right tail in contrast to visible orders which seem to effect a relatively concave power-law far-right tail. Moreover, the average market impact curves are volatile after the introduction of hidden orders especially towards the far-left of the impact curves. The observation suggests that hidden orders posses a material effect on the structure of the market impact of orders; which, translates into hidden orders, and have a significant influence over prices of stocks. Based on the observations of Figure 1 and Figure A.21 we comment that indeed hidden orders affect the structure of the average market impact curves.

The observation that hidden orders affect the average market impact curves was of special interest to us. Agents do not necessarily know at least the sizes of hidden orders in the neighborhood of the duration of the trading period, unless, agents have insider information. A trader that seeks to practice cost-effective trading that minimizes market impact would be concerned about the arrival of hidden orders in the market. We present machine learning solutions to revealing what the sizes of hidden orders are. Tables 1, 2, 11, 16 & 21 indicate exceptional performance on average for machine learning on the estimation of the normalized sizes of hidden orders. We record large R^2 and small $\hat{\mathrm{MSE}}$ in the statistics that are disclosed in the tables. Hence machine learning can, with limitations, accurately predict the sizes of hidden orders. Not only can we uncover the sizes of hidden orders, but we could calculate the average market impact curves of visible orders and machine estimated hidden orders that are close to the average market impact curves of visible orders and true hidden orders. Figures 11 - 12 shows that the mean of the average market impact curves of visible orders and machine estimated hidden orders approximates the mean of the average market impact curves of visible orders and true hidden orders well. Tables 1, 2, 11, 16 & 21 and Figures 11 - 12 support the title of our research.

We brief the unfamiliar researcher on how the results of Figures 11 - 12 were derived. The average market impact curve of visible orders and true hidden orders was deduced by applying the approach of Lillo *et al.* (2003) on the data of visible orders and true hidden orders of individual firms. We averaged the individual average market impact curves to determine the mean curve of the individual average market impact curves. The average market impact curve of visible orders and estimated hidden orders was deduced by applying the approach of Lillo *et al.* (2003) on the data of visible orders and machine estimated hidden orders of individual firms using the techniques of Sections 2.1 - 2.4. We averaged the individual average market impact curves to determine the mean curve of the individual average market impact curves.

The confidence intervals were derived from calculating the varying standard deviation of the average price shifts of the individual firms in specified normalized volume bins. The reason we allowed the standard deviation to vary throughout was to introduce some degree of uncertainty on the calculations such that we could derive a confidence interval about how we interpret our results from here on. We multiplied the critical value of 90% with varying standard deviation figures and divided the product by the square root of the bin sample size. The results assist us to turn casual comments into contextual comments that are substantiated. We observe that as the normalized volume increases over the mean of the average normalized volume, the confidence intervals become narrow. The observation suggests that with 90% confidence we are more and more confident that as normalized volume increases above the mean of the average normalized volumes, the market impact increases as a power law. The confidence intervals are much wider for normalized volumes less than the mean of the average normalized volume. The observation suggests that we could argue that market impact is a steadily decreasing function of normalized volume for orders of size less and less than the mean of the average normalized volumes. However, one could also argue that market impact is flat for orders of size strictly less than the mean of the average normalized volumes because the wider interval suggests that we are less confident in the interpretation of our results.

We warn researchers that artificial intelligence may not always be a reliable tool to apply. We examined the performance of our Generalized Linear Models (GLM) to determine if we could find common features that could be used to predict the sizes of hidden orders across a variety of firms. According to Figure 6, the mid quote price before, liquidity and KL divergence volume are not useful to the linear learning algorithms. Meanwhile, the turnover ratio contributes significantly. The observation suggests that in an ideal firm there exists a correlation between the volume of shares executed and the turnover ratio. Large-capitalization firm: CPI and small-capitalization firm: ADH have similar feature contribution effects. Figure 6 shows that although there are quite several features that are exploited by the GLM to predict the sizes of hidden orders, there exists no set of market features that are consistently predictive of the sizes of hidden orders across different stocks. We held the computing algorithms constant and proceeded to seek answers to our research questions from the BOVESPA, MOEX, NSE and SSE markets.

3.2. Brazil (BOVESPA)

We observe that small average daily value traded firms and large average daily value traded firms tend to conflate (Figure 2). The observation suggests that market capitalization of firms may not necessarily be an influential factor that separates the cost structures of small firms from large firms. Other notable observations of Figure 2 are that the average price shift lies roughly in between the range $[10^{-4.1}, 10^{+0.3}]$. Figure A.22 shows that different from the JSE, hidden orders in the BOVESPA market seem to affect the average market impact curves for order sizes of less than the mean of the average normalized volumes. Tables 3, 4, 12, 17 & 22 indicate that machine learning still has relevance on the estimation of hidden orders. Moreover, we observe in Figures 13 - 14 that machine learning produces the mean curve of the average price impact curves of visible orders and machine estimated hidden orders lie slightly below the mean curve of the average price impact curves of visible orders

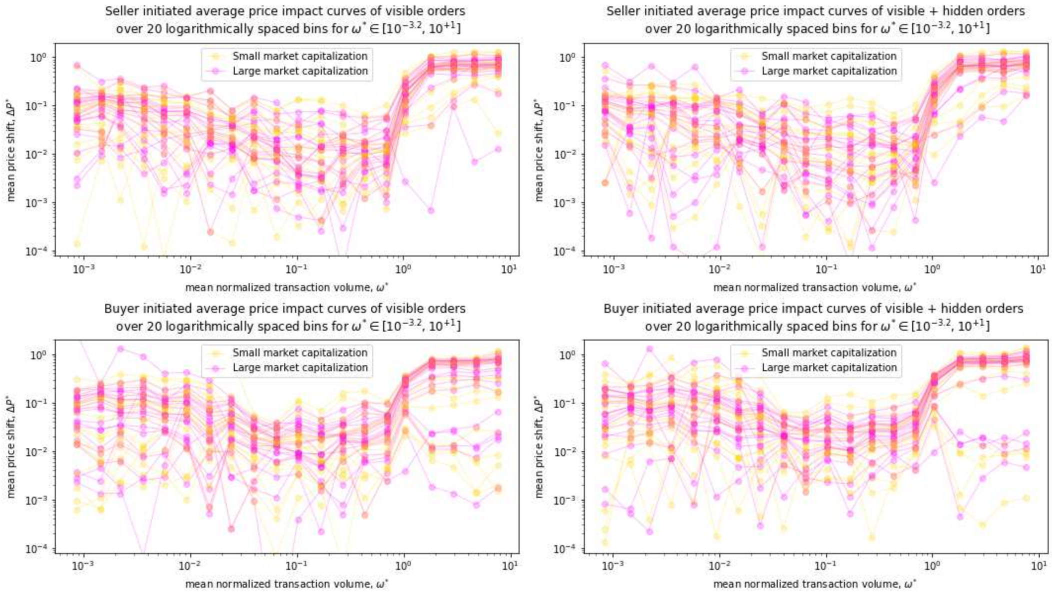

Fig. 2. Price impact curves of single trades of 42 liquid firms. The curves are computed using the analytical literature approach. (Lime Curves) represent firms with the average daily value traded larger than the median average daily value traded. (Deeppink Curves) represent firms with the average daily value traded smaller than the median average daily value traded.The analysis is period from 28-09-2018 to 28-06-2019.

Bolsa de Valores de Sao Paulo (BOVESPA). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

and true hidden orders. The observation suggests that machine learning would still be relevant for the prediction of the average market impact of hidden orders. On average we tend to perform very well for orders of sizes less than the mean normalized order size and fairly well for orders of size greater than the mean normalized order size.

In Figures 13 - 14, we observe that the average curves with true hidden orders separate significantly from the average curves with estimated hidden orders. However, the trend is the same. We argue that the BOVESPA market may not have a significant number of hidden orders that are large. According to Figure 7 price, momentum and KL divergence volume are not useful to the linear learning algorithms. Meanwhile, the KL divergence counts contribute significantly. The observation suggests that in an ideal firm there exists a correlation between the volume of shares executed and the KL divergence counts. Large-capitalization firm: ITSA4 and small-capitalization firm: AZE4 have similar feature contribution effects. Similar to the JSE there exists no set of market features that are consistently predictive of the sizes of hidden orders across different stocks as supported by Figure 7.

3.3. Russia (MOEX)

We observe that small average daily value traded firms tend to dominate the top of the overall curves for seller initiated trading; meanwhile, large average daily value traded firms tend to dominate the top of the overall curves for buyer initiated trading as indicated in Figure 3. We reshuffled the plots to enable our program to plot the results randomly to retrieve what occurs in the MOEX market. We observe that for seller initiated trading large firms tend to cluster at the bottom; meanwhile, small firms tend to cluster at the top. However, the opposite occurs in buyer initiated trading where large firms tend to cluster at the top and small firms tend to cluster at the bottom. If our observation holds true, then we could say that there exists a clear arbitrage in the MOEX market. Seller initiated traders of large firms do not have to worry about market impact because the buyer initiated traders of large firms are paying excessive market costs for trading. Similarly, buyer initiated traders of small firms do not worry much because seller

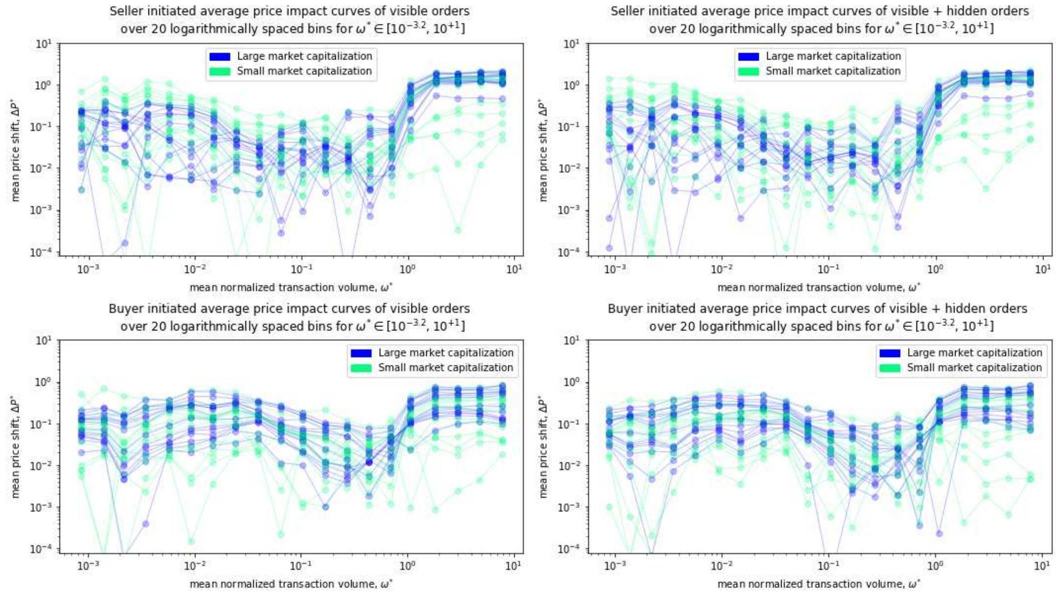

Fig. 3. Price impact curves of single trades of 30 liquid firms. The curves are computed using the analytical literature approach. (Teal Curves) represent firms with the average daily value traded larger than the median average daily value traded. (Thistle Curves) represent firms with the average daily value traded smaller than the median average daily value traded. The analysis is period from 28-09-2018 to 28-06-2019.

 Moscow Exchange (MOEX). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

initiated traders of small firms are paying market costs on their behalf. We are not in a position to comment further on these observations, since we do not have further information. However, we speculate that MOEX market rules and regulations may be possible causes of unique observations. The role of market capitalization on the structure of impact curves seems to be an open question because numerous issues prohibit us from comparing the analysis of the MOEX data with the analyses of the data from other markets. Similar notable observations of Figure 3 are that the average price shift lies roughly in between the range $[10^{-4.1}, 10^{+1}]$.

Figure A.23 shows that similar to the BOVESPA market, hidden orders in the MOEX seem to introduce the variance effect on the average market impact curves for order sizes of less than the mean of the average normalized order sizes. Tables 5, 6, 13, 18 & 23 indicate that although machine learning still shows relevance on the estimation of hidden orders, there is generally a poor contribution of features on the estimation of hidden orders. Figure 8 indicates that most of the features did not contribute to linear algorithms towards the estimation of hidden orders. Thus our machine learning results for the MOEX market may be biased to the market and the data used. Perhaps several issues needed to be brought into consideration when we trained our machine learning techniques. Moreover, we observe in Figures 15 - 16 that machine learning produces the mean curve of the average price impact curves of visible orders and machine estimated hidden orders lie slightly below the mean curve of the average price impact curves of visible orders and true hidden orders. Except for support vector machines which perform relatively poorly. The observations suggest that machine learning would still be relevant for the prediction of the average market impact of hidden orders with strict limitations though. On average we tend to perform very well for orders of sizes close to the mean normalized order size and fairly well for orders of size greater than the mean normalized order size. We comment that the poor learning of our machine learning techniques might have been caused by the fact that there were fewer hidden orders in the MOEX data. The effect of fewer hidden has resulted in an imbalance in our machine learning techniques which limited their predictive capability. We also speculate that the MOEX order placements adopt a unique structure that hinders our machine learning.

On average we tend to under perform due to poor performance of machine learning on the majority of stocks. According to Figure 8 the spread before, mid quote price before, liquidity, volatility, and KL divergence volume are not useful to the linear learning algorithms. Meanwhile, the signed omega contributes significantly. The observation suggests that in an ideal firm there exists a correlation between the volume of shares executed and the signed omega. Large-capitalization firm: NLMK and small-capitalization firm: BSPB have similar feature contribution effects. Similar to the JSE there exists no set of market features that are consistently predictive of the sizes of hidden orders across different stocks as supported by Figure 8.

3.4. *India (NSE)*

Observations that are similar to the BOVESPA market are derived as indicated in Figure 4. Other notable observations of Figure 2 are that the average price shift lies roughly in between the range $[10^{-4.5}, 10^{+1}]$. We cannot tell as to what role market capitalization plays, because of the chaotic behavior of the average market impact curves of large average daily value traded firms and small average daily value traded firms. Thus, market capitalization may not possess a significant role on the market costs of small firms and large firms. We furthermore observe that buyer initiated trading market impact curves have a power-law far-right tail in contrast to seller initiated curves which have a mixture of structures (Figure 4). Figure A.24 shows that similar to BOVESPA and MOEX, hidden orders in the NSE seem to have a volatile material effect on the impact curves for orders of size less than the average normalized order size.

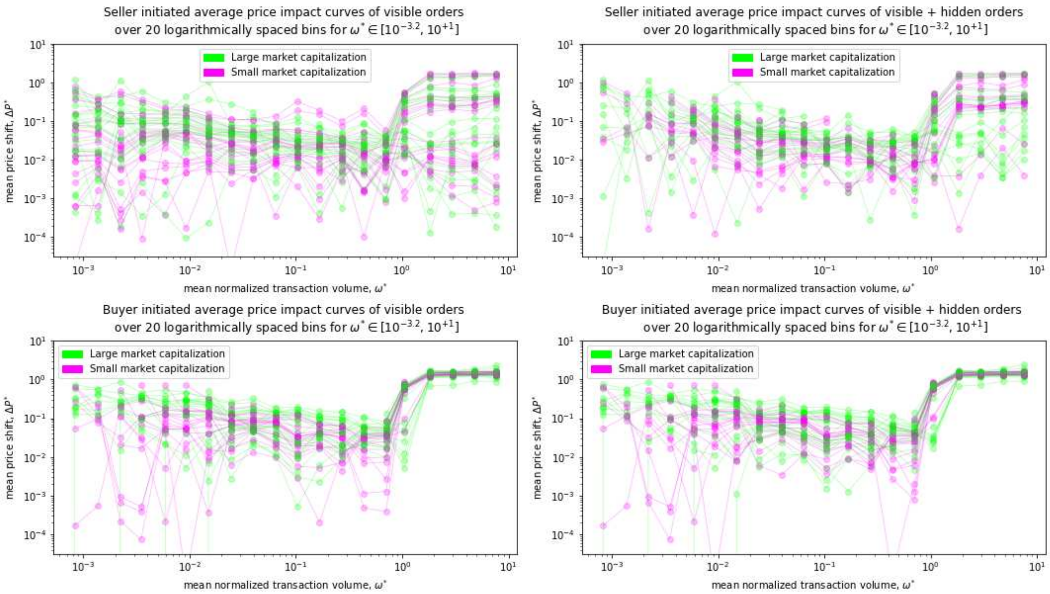

Fig. 4. Price impact curves of single trades of 43 liquid firms. The curves are computed using the analytical literature approach. (Lime Curves) represent firms with the average daily value traded larger than the median average daily value traded. (Magenta Curves) represent firms with the average daily value traded smaller than the median average daily value traded. The analysis is period from 28-09-2018 to 28-06-2019.

National Stock Exchange of India (NSE). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

Tables 7, 8, 14, 19 & 24 indicate that machine learning still has relevance on the estimation of hidden orders. Moreover, we observe in Figures 17 - 18 that generalized linear models and artificial neural networks produce the mean curve of the average price impact curves of visible orders and machine estimated hidden orders that intercept the mean curve of the average price impact curves of visible orders and true hidden orders. Meanwhile, support vector machines and random forests produce the mean curve of the average price impact curves of visible orders and machine estimated hidden orders that lies below the mean curve of the average price impact curves of visible orders and true hidden orders. The very interesting observation is that the average market impact curve of visible orders and machine estimated hidden orders are linear irrespective of the technique used. The observation suggests that order placement may be following a linear function in the NSE. Machine learning would still be relevant for the prediction of the average market impact of hidden orders. On average we tend to perform very well for orders of all sizes.

According to Figure 9, mid quote price before, liquidity, volatility and KL divergence volume are not useful to the linear learning algorithms. Meanwhile, the bid size contributes significantly. The observation suggests that in an ideal firm there exist a correlation between the volume of shares executed and the bid size. Large-capitalization firm: VAKR and small-capitalization firm: SUBR have similar feature contribution effects. Similar to the JSE there exist no set of market features that are consistently predictive of the sizes of hidden orders across different stocks as supported by Figure 9.

3.5. China (SSE)

We observe that small average daily value traded firms and large average daily value traded firms appear on a single scale for both sellers initiated trading and buyer initiated trading as indicated in Figure 5. The observation suggests that market capitalization has no role in the structure of the market impact curves. Therefore the costs of trading a small firm and the costs of trading a large firm are the same.

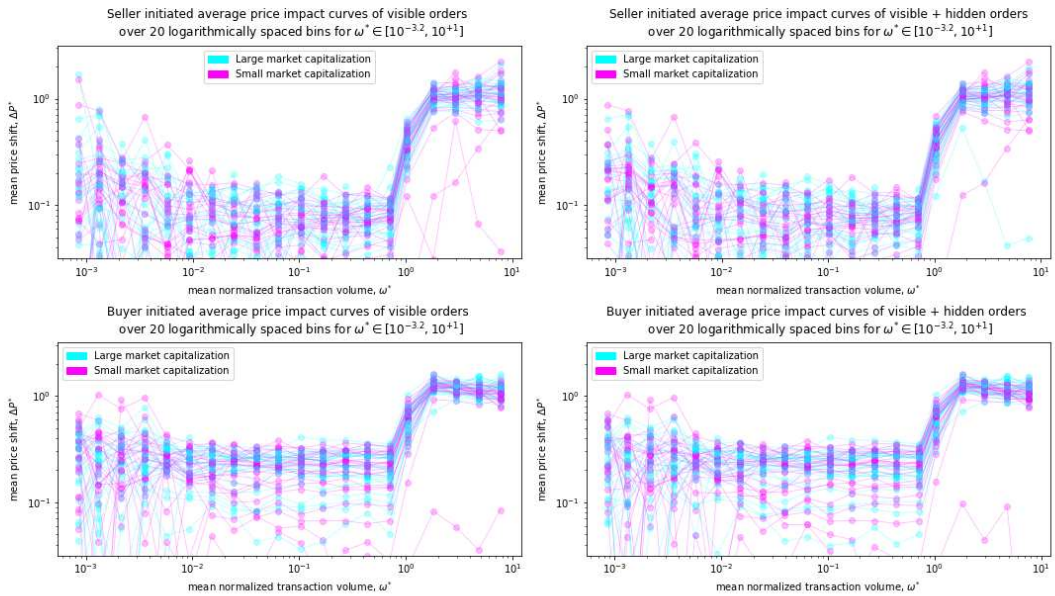

Fig. 5. Price impact curves of single trades of 54 liquid firms. The curves are computed using the analytical literature approach. (Cyan Curves) represent firms with the average daily value traded larger than the median average daily value traded. (Magenta Curves) represent firms with the average daily value traded smaller than the median average daily value traded. The analysis is period from 28-09-2018 to 28-06-2019.

Shanghai Stock Exchange (SSE). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

Harvey *et al.* (2016) has reported similar results. Other notable observations of Figure 2 are that the average price shift lies roughly in between the range $[10^{-1}, 10^{+0.5}]$. Figure A.25 shows that hidden orders in the SSE seem to have no material effect on the impact curves at all. The observation suggests that hidden orders may not necessarily be large or the SSE market is highly liquid enabling it to absorb significantly large orders at minimal market costs. Tables 9, 10, 15, 20 & 25 indicate that machine learning still has relevance on the estimation of hidden orders. We observe in Figure 19 - 20 that machine learning produces the mean curve of the average price impact curve of visible and machine estimated hidden orders lies below the mean curve of the average price impact curve of visible and true hidden orders. On average, we tend to perform fairly well due to the adequate performance of our machine learning techniques on the majority of stocks.

According to Figure 10 the bid price, liquidity, volatility, and KL divergence volume are not useful to linear learning algorithms. Meanwhile, the bid size contributes significantly. The observation suggests that in an ideal firm there exists a correlation between the volume of shares executed and the bid size. The generalized linear models performed poorly on the two firms 600889 and 600835. Large-capitalization firm: 601333 and small-capitalization firm: 600893 have similar feature contribution effects. Similar to the JSE there exists no set of market features that are consistently predictive of the sizes of hidden orders across different stocks as supported by Figure 10. Unlike in the other markets, the average market impact curve of visible orders and machine estimated hidden orders appear outside the original scale of $[10^{-1}, 10^{+0.5}]$. This has lead to an increase in scale range to $[10^{-4.5}, 10^{+0.5}]$. The observation suggests that either it is difficult to train machine learning on the SSE data or our analytical approach of Lillo *et al.* (2003) tends to overestimate the average market impact.

4. Conclusions

Market capitalization does not necessarily affect the average market impact of firms. The role of market capitalization on the average market impact curves differ across markets. The degree to which hidden orders affect the average market impact of firms depends on the liquidity of the market, and possibly market depth. Machine learning can, with limitations and varying degrees of exploitation of the features, accurately predict the sizes of hidden orders. It is ideally possible to construct an average market impact curve of machine learning estimated hidden orders, that adequately approximates the average market impact curve of true hidden orders. Future research prospects would entail an investigation as to why some of the average market impact curves, in which hidden orders are estimated via machine learning, do not converge close to the average market impact curves of true hidden orders.

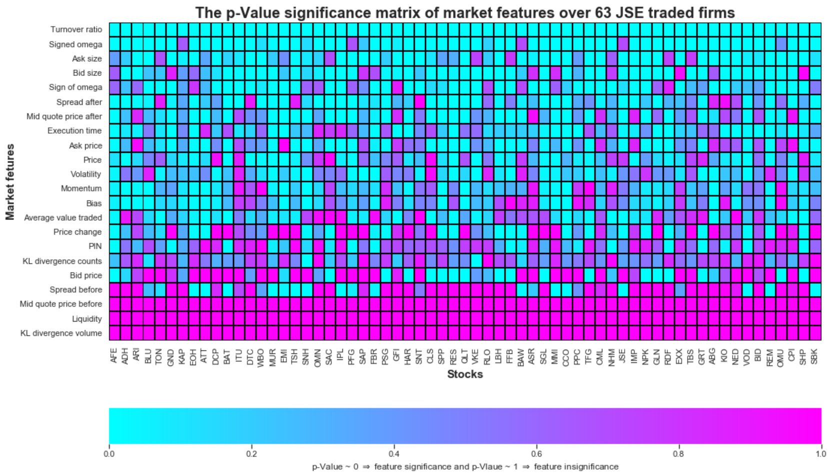

Fig. 6. The p-value significance matrix of 63 JSE traded firms. The firms are ordered according to average daily value traded from small average daily value traded firms (left) to large average daily value traded firms (right). The p-Value reveals how significant a specified market feature is, with regard to the estimation of normalized transaction sizes using a generalized linear model. p-Value has a common scale of between 0 and 1, and that allows a cross examination of features across firms to establish patterns.

Johannesburg Stock Exchange (JSE). *Approach* : Lillo *et al.* (2003) & Seber *et al.* (1977)

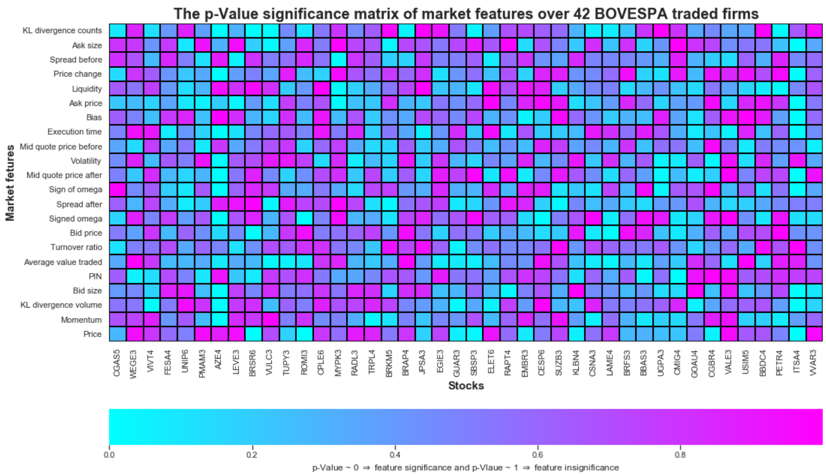

Fig. 7. The p-value significance matrix of 42 BOVESPA traded firms. The firms are ordered according to average daily value traded from small average daily value traded firms (left) to large average daily value traded firms (right). The p-Value reveals how significant a specified market feature is, with regard to the estimation of normalized transaction sizes using a generalized linear model. p-Value has a common scale of between 0 and 1, and that allows a cross examination of features across firms to establish patterns.

Bolsa de Valores de Sao Paulo (BOVESPA). *Approach* : Lillo *et al.* (2003) & Seber *et al.* (1977)

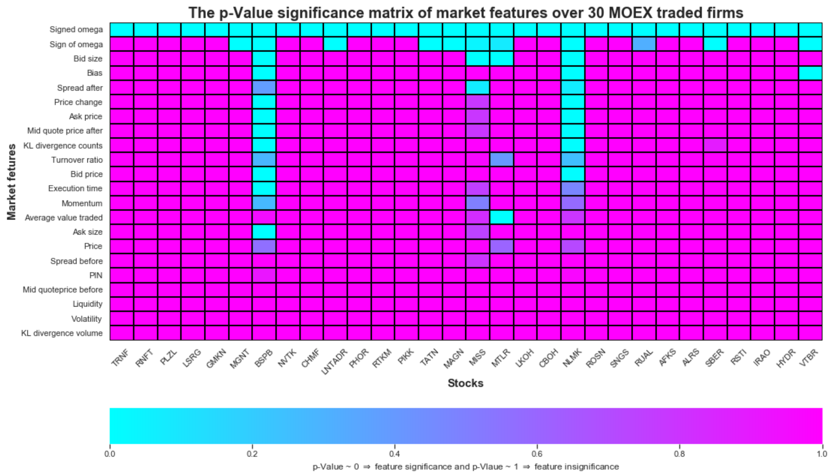

Fig. 8. The p-value significance matrix of 30 MOEX traded firms. The firms are ordered according to average daily value traded from small average daily value traded firms (left) to large average daily value traded firms (right). The p-Value reveals how significant a specified market feature is, with regard to the estimation of normalized transaction sizes using a generalized linear model. p-Value has a common scale of between 0 and 1, and that allows a cross examination of features across firms to establish patterns.

Moscow Exchange (MOEX). *Approach* : Lillo *et al.* (2003) & Seber *et al.* (1977)

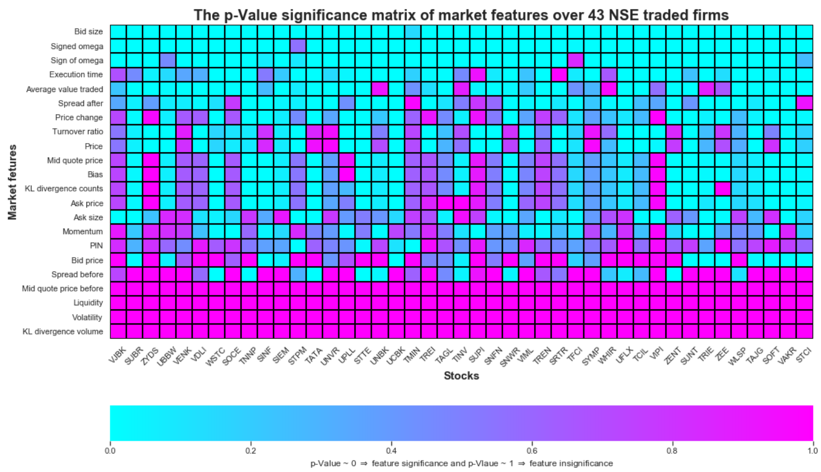

Fig. 9. The p-value significance matrix of 43 NSE traded firms. The firms are ordered according to average daily value traded from small average daily value traded firms (left) to large average daily value traded firms (right). The p-Value reveals how significant a specified market feature is, with regard to the estimation of normalized transaction sizes using a generalized linear model. p-Value has a common scale of between 0 and 1, and that allows a cross examination of features across firms to establish patterns.

National Stock Exchange of India (NSE). *Approach* : Lillo *et al.* (2003) & Seber *et al.* (1977)

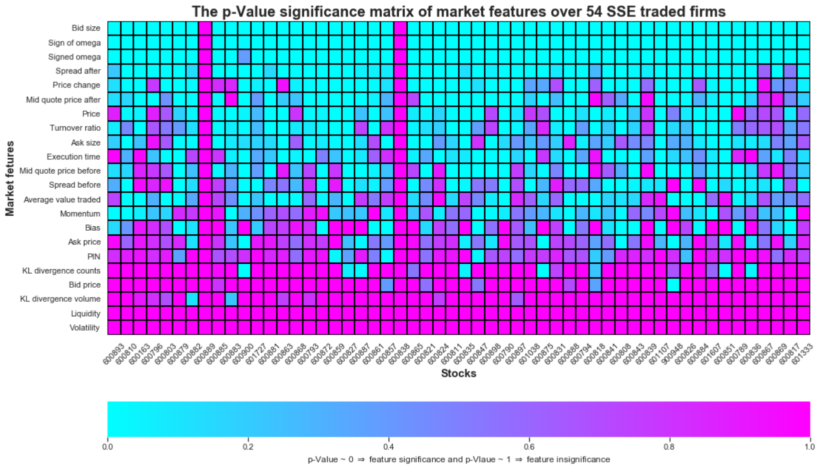

Fig. 10. The p-value significance matrix of 54 SSE traded firms. The firms are ordered according to average daily value traded from small average daily value traded firms (left) to large average daily value traded firms (right). The p-Value reveals how significant a specified market feature is, with regard to the estimation of normalized transaction sizes using a generalized linear model. p-Value has a common scale of between 0 and 1, and that allows a cross examination of features across firms to establish patterns.

Shanghai Stock Exchange (SSE). *Approach* : Lillo *et al.* (2003) & Seber *et al.* (1977)

References

Alfonsi, A., Schied, A., Slynko, A., 2009. Order book resilience, price manipulation, and positive portfolio problem. SIAM Journal on Financial Mathematics. https://ssrn.com/abstract=1498514

Alfonsi, A., Schied, A., 2010. Optimal trade execution and absence of price manipulations in limit order book models. SIAM Journal on Financial Mathematics, 7/090762786ff. ffhal-00397652v3f

Amihud, Y., Mendelson, H., 1986. Asset pricing and the bid-ask spread. *Journal Financial Economics*, (17): 223 - 249. https://www.sciencedirect.com/science/article/pii/0304405

Bouchaud, J.P., 2017. Price Impact, *Capital Fund Management*, Encyclopedia of Quantitative Finance, R. Cont Edt.; [q-fin.TR], arXiv: 0903.24.281. John Wiley Sons Ltd. https://arxiv.org/abs/09

Breiman, L., 2001. Random Forests, *Machine learning*, Volume. 45, page. 5–32, https://link.springer.com/article/10.1023/A:1010933404324

Easely, D., Keifer, N.M., O'Hara, M., 2002. Is information risk a determinant of asset returns? *Journal of Finance*, 57(5):2185-2221. https://onlinelibrary.wiley.com/doi/full/10.1111/1540-6261.

Fan, R.E., Chen P.H., Lin C.J., 2005. Working Set Selection Using Second Order Information for Training Support Vector Machines. The Journal of Machine Learning Research, No: 6: 1871–1918.

Farmer, J.D., Lillo, F., 2004. On the origin of power-law tails in price fluctuations. Quantitative Finance, 4: C7–C11. [Taylor Francis Online], https://pdfs.semanticscholar.org/1c6a/4027739960d02b8ac

Gatheral, J., Schied, A., Slynko, A., 2010. Transient linear price impact and Fredholm integral equations. Preprint available at https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-99

Gatheral, J., 2011. No-Dynamic-Arbitrage and Market Impact. To appear in Quantitative Finance, Preprint available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.

Glosten, L.R., Milgrom, P.R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. J. Financ. Econ., 14: 71–100. https://econpapers.repec.org/article/ee

Hagan, M.T., Menhaj, M., 1994. "Training feedforward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, Vol. 5, No. 6, pp. 989–993. https://www.sciencedirect.com

Harvey, M., Hendricks, D., Gebbie, T., Wilcox, D., 2016. Deviations in expected price impact for small transaction volumes under fee restructuring, Elsevier B.V. Physica A, 471, pp 416-426.1602.049

Huang, T.M., Kecman, V., Kopriva, I., 2006. Kernel-Based Algorithms for Mining Huge Data Sets: Supervised, Semi-Supervised, and Unsupervised Learning. Springer, New York.97835403168

Huberman, G., Stanzl, W., 2004. Price Manipulation and Quasi-Arbitrage, Econometrica, Vol.74, pp 1247-1276. https://econpapers.repec.org/article/ecmemetrp/v3a723ay3a

Jegadeesh, N., Sheridan T., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency, Journal of Finance 48, 65-91. https://econpapers.repec.org/article/blajfinan/v3a483

Klock, F., Schied, A., Sun, Y., 2011. Existence and absence of price manipulation in a market impact model with a dark pool. Available at SSRN 1785409. https://www.researchgate.net/publication/2281201

Kyle, AS., 1985. Continuous auctions and insider trading. Econometrica, 53(6): 1315–1336. [Crossref], [Web of Science ®], [Google Scholar] https://www.rhsmith.umd.edu/files/Documents/Centers/CFP

Kullback, S., Leibler, R.A., 1951. "On information and sufficiency". Annals of Mathematical Statistics. 22 (1): 79–86. doi:10.121.http://mathfaculty.fullerton.edu/sbehseta/Kullback.

Lee, C.M.C., Ready, M.J., 1991. Inferring Trades Direction from Intraday Data, The Journal of Finance, Vol.46, No.2, pp 733-746. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.

Lillo, F., Mantegna, R., Farmer, J.D., 2003. Master Curve for Price-Impact Function, Nature, Vol 421. https://www.nature.com/articles/421129a?message=remove

Marquardt, D., 1963. "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," SIAM Journal on Applied Mathematics, Vol. 11, No. pp. 431–441.https://www.scirp.org/Reference=1409246

Miccichè, S., Bonanno, G., Lillo, F., and Mantegna, R.N., 2002. Physica A 314 756. https://link.springer.com/chapter/10.1007/978-4-431-53947-6-4. DOI: 10.1016/S0378-4371(02)01187-1./0202527

Moro, E., Vicente, J., Moyano, L.G., Gerig, A., Farmer, J.D., Vaglica, G., Lillo, F., Mantegna, R.N., 2009. Market impact and trading profile of hidden orders in stock markets, Physical R E 80, 066102.

Nguyen, D., Widrow, B., 1990. "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," Proc. of the Int. J. Conf. on Neural Networks, (3), pp. 21–26.

O'Hara, M., 1995. Market Microstructure Theory, Blackwell, Oxford, ISBN 1-55786-443-8.The Review of Financial Studies, Vol. 8, Is. 4, pg 1235, https://academic.oup.com/rfs/article/8/4/1235/1581867

Potters, M., Bouchaud, J.P., 2003. More Statistical Properties of order books and price impact, Physica, A,324, 133. Statistical Mechanics and its Applications. http://www.scirp.org/Reference=808166

Seber, G.A.F., 1977. Linear Regression Analysis. New York: Wiley. 2nd edition. https://www.academia.edu/32085934/Linear-Regression-Analyisis-2nd-edition-George-A.F.Seber-Alan-J.Lee.04710196

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York. https://statisticalsupportandresearch.files.wordpress.com/2017/05/vladimir-vapnik-the-nature-of-statistical-learning-springer.

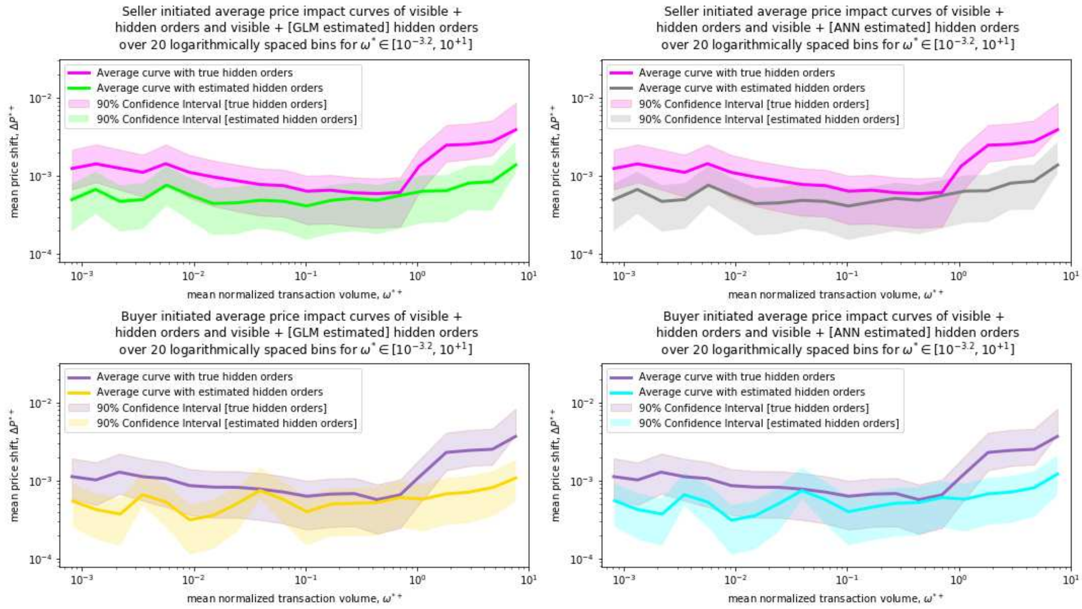

Fig. 11. The mean price impact curves and the 90% confidence interval from the mean of single trades of 63 liquid firms. The sizes of hidden orders are estimated through generalized linear model (left plots), and artificial neural networks (right plots).

Johannesburg Stock Exchange (JSE). *Approach* : Sebber *et al.* (1977) & Nguyen *et al.* (1990)

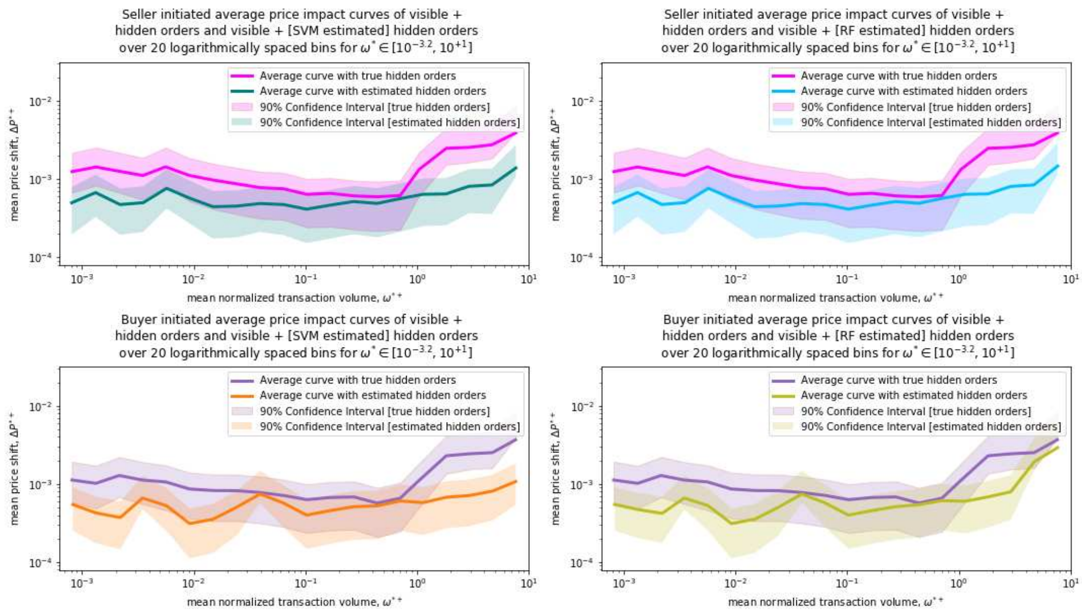

Fig. 12. The mean price impact curves and the 90% confidence interval from the mean of single trades of 63 liquid firms. The sizes of hidden orders are estimated through support vector machines (left plots), and random forests (right plots).

Johannesburg Stock Exchange (JSE). *Approach* : Vapnik *et al.* (1995) & Breiman *et al.* (2001)

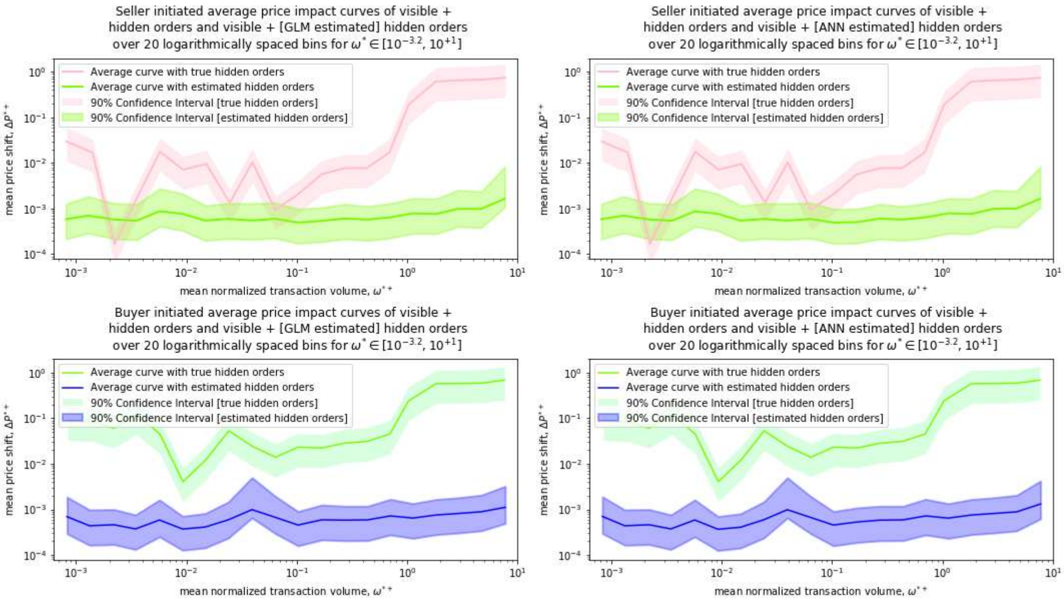

Fig. 13. The mean price impact curves and the 90% confidence interval from the mean of single trades of 42 liquid firms. The sizes of hidden orders are estimated through generalized linear model (left plots), and artificial neural networks (right plots).

Bolsa de Valores de Sao Paulo (BOVESPA). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

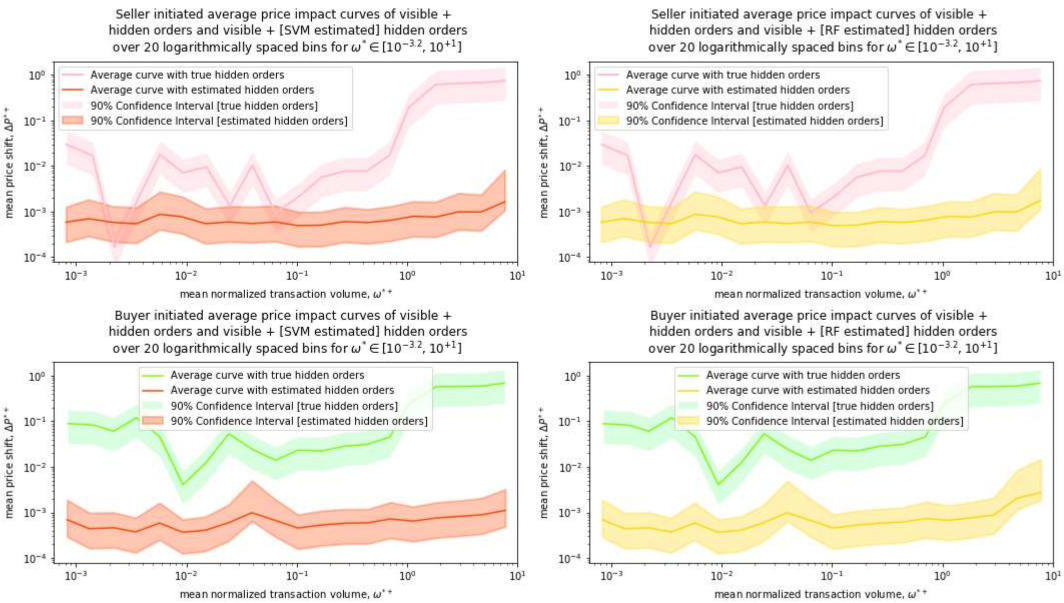

Fig. 14. The mean price impact curves and the 90% confidence interval from the mean of single trades of 42 liquid firms. The sizes of hidden orders are estimated through support vector machines (left plots), and random forests (right plots).

Bolsa de Valores de Sao Paulo (BOVESPA). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

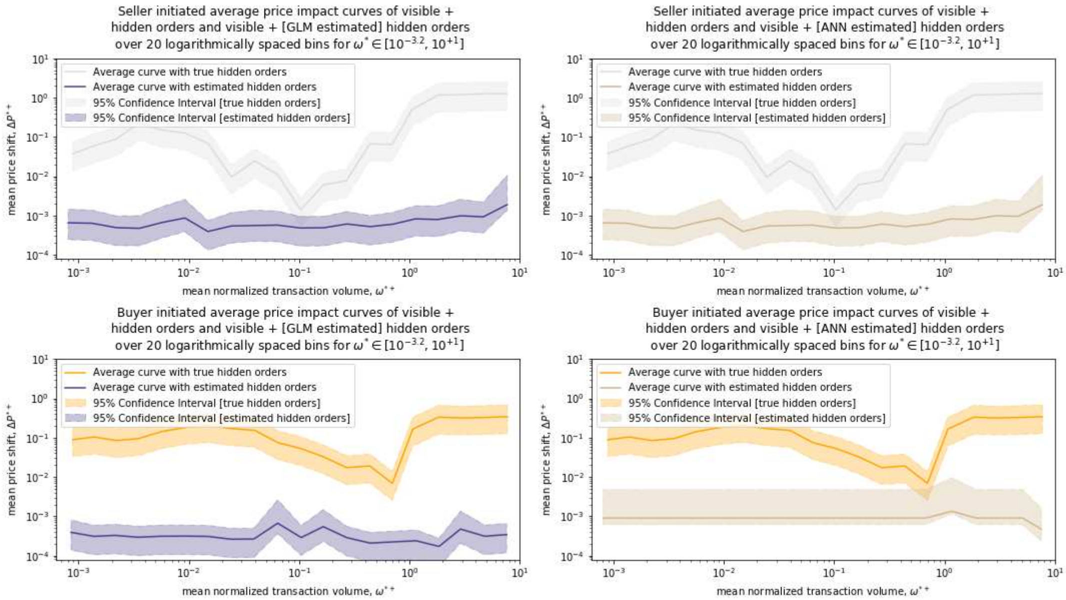

Fig. 15. The mean price impact curves and the 90% confidence interval from the mean of single trades of 30 liquid firms. The sizes of hidden orders are estimated through generalized linear model (left plots), and artificial neural networks (right plots).

Moscow Exchange (MOEX). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

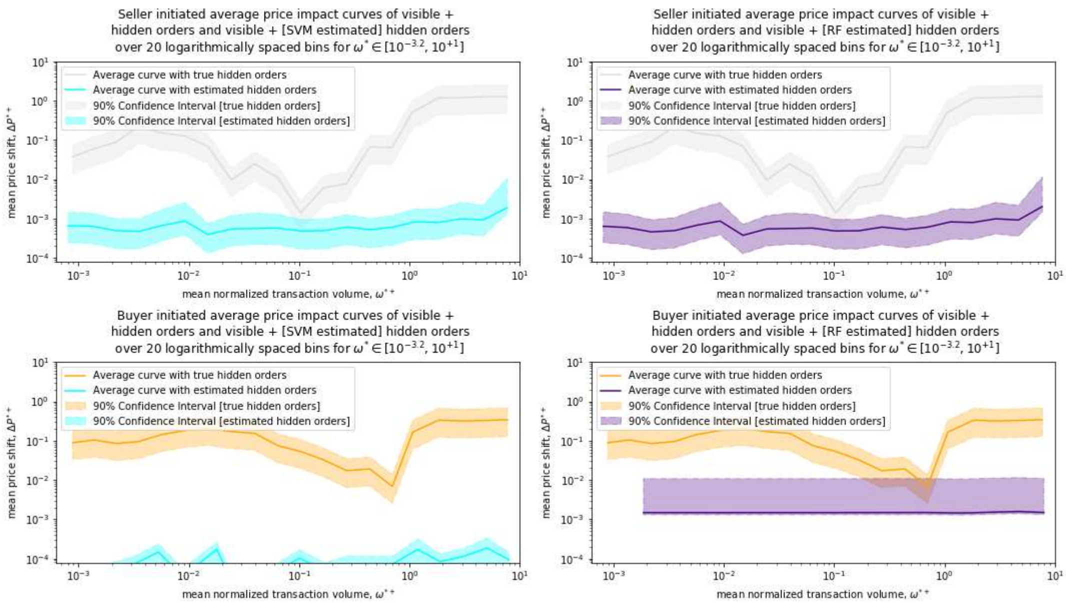

Fig. 16. The mean price impact curves and the 90% confidence interval from the mean of single trades of 30 liquid firms. The sizes of hidden orders are estimated through support vector machines (left plots), and random forests (right plots).

Moscow Exchange (MOEX). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

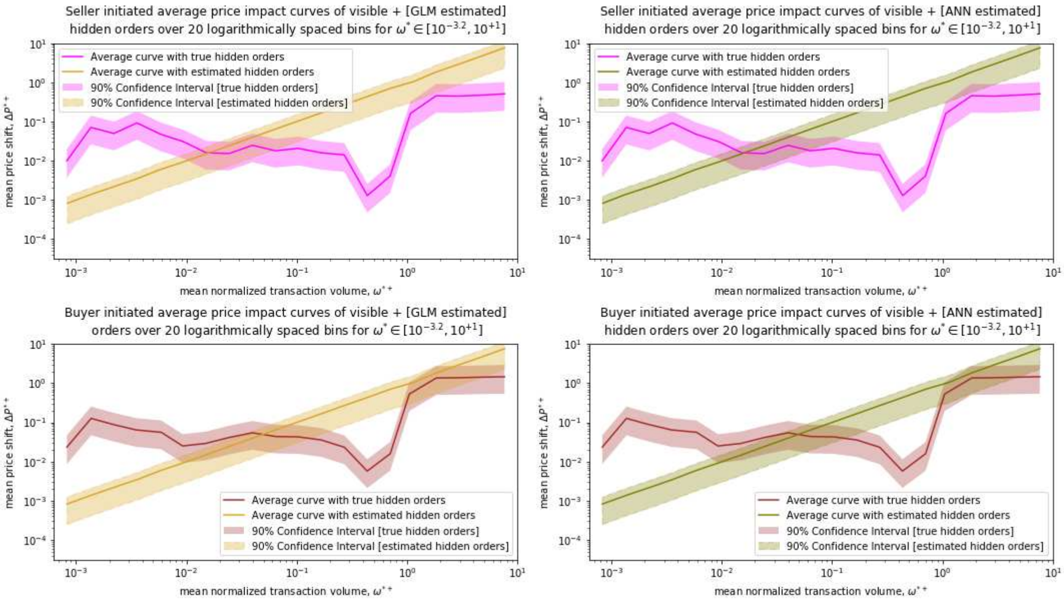

Fig. 17. The mean price impact curves and the 90% confidence interval from the mean of single trades of 43 liquid firms. The sizes of hidden orders are estimated through generalized linear model (left plots), and artificial neural networks (right plots).

National Stock Exchange of India (NSE). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

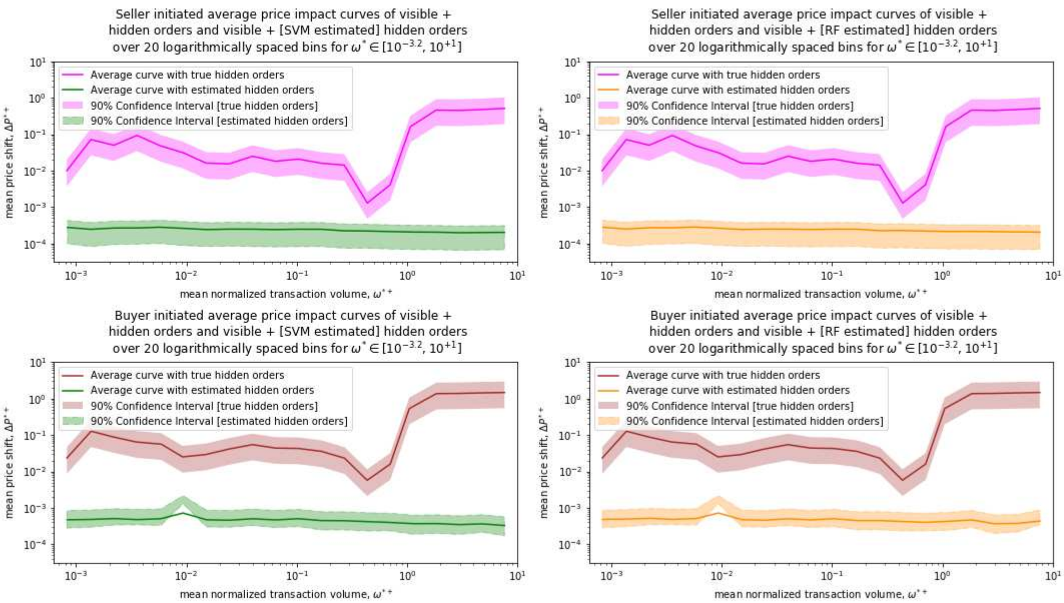

Fig. 18. The mean price impact curves and the 90% confidence interval from the mean of single trades of 43 liquid firms. The sizes of hidden orders are estimated through support vector machines (left plots), and random forests (right plots).

National Stock Exchange of India (NSE). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

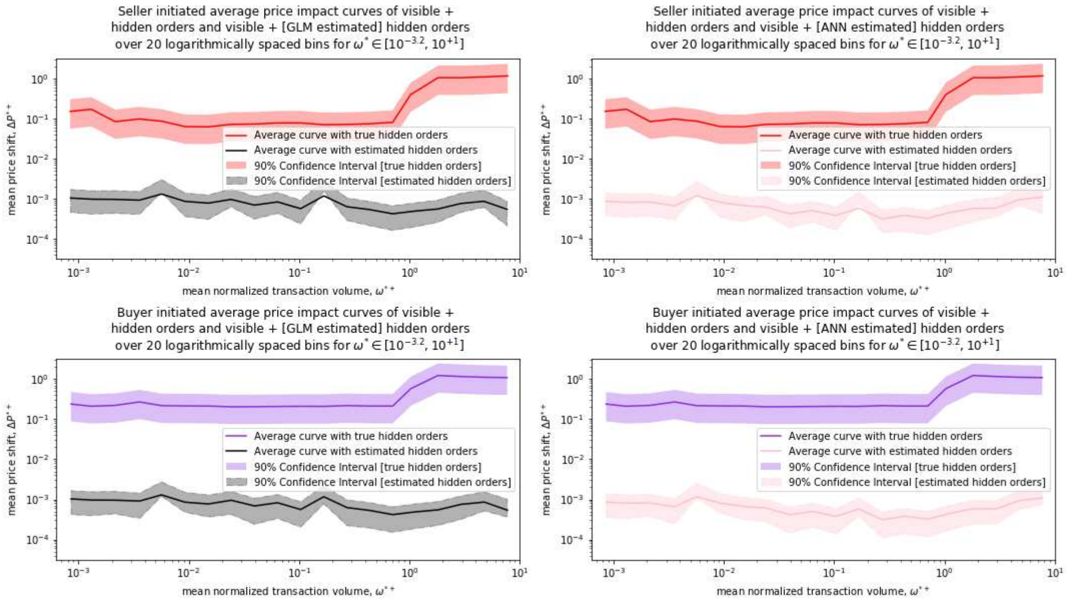

Fig. 19. The mean price impact curves and the 90% confidence interval from the mean of single trades of 54 liquid firms. The sizes of hidden orders are estimated through generalized linear model (left plots), and artificial neural networks (right plots).

Shanghai Stock Exchange (SSE). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

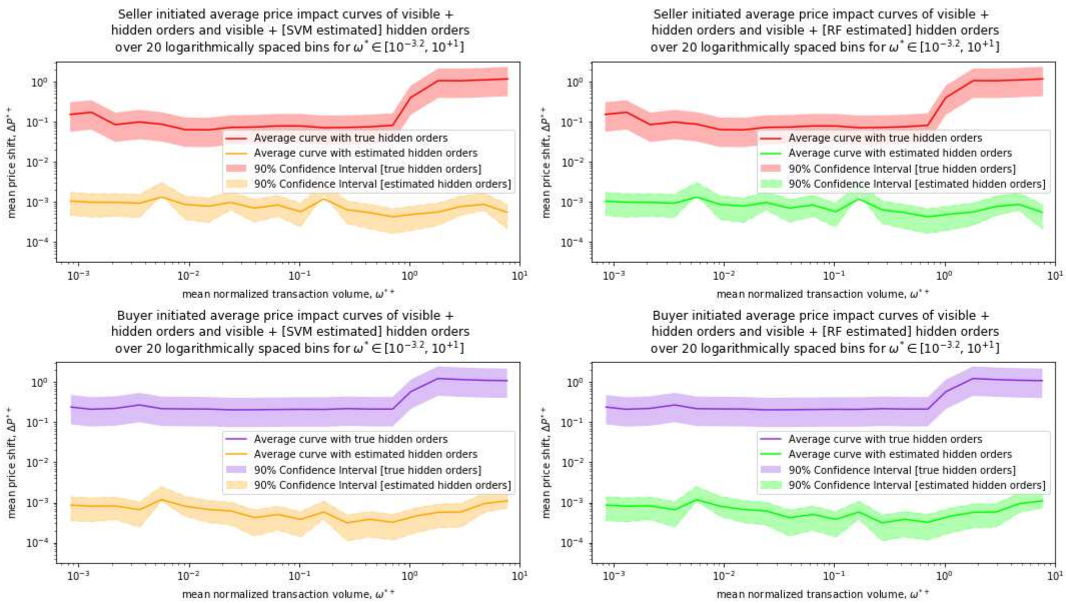

Fig. 20. The mean price impact curves and the 90% confidence interval from the mean of single trades of 54 liquid firms. The sizes of hidden orders are estimated through support vector machines (left plots), and random forests (right plots).

Shanghai Stock Exchange (SSE). *Approach* : Lillo *et al.* (2003) & Harvey *et al.* (2016)

Table 1. Johannesburg Stock Exchange (JSE). GLM parameters of the range and the average across top 63 highly liquid stocks.

Feature	Feature description	Metric	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	range	[-308.16, 147.82]	[4.40, 255.08]	[-6.64, 4.02]	[4.2E-11, 0.98]
		mean	-2.1223	30.6957	0.1792	0.3282
x_{i1}	Price	range	[-0.11 , 0.09]	[0.00 , 0.15]	[-5.45, 3.53]	[5.9E-08, 0.97]
		mean	-0.0056	0.0157	-0.1241	0.2963
x_{i2}	Bid price	range	[-1.69, 3.47]	[0, 1.64]	[-7.09, 32.04]	[6E-173,0.96]
		mean	0.0905	0.1423	1.7529	0.2322
x_{i3}	Bid size	range	[-0.00, 3.47]	[4.5E-07, 1.64]	[-4.44, 32.04]	[0 , 0.96]
		mean	0.0001	3.5E-05	4.5990	0.1748
x_{i4}	Ask price	range	[-9.12, 3.44]	[0 , 3.44]	[-35.71, 16.14]	[2E-208, 0.97]
		mean	-0.1049	0.3886	-0.4519	0.2599
x_{i5}	Ask size	range	[-0.00, 0.00]	[2.8E-07, 0.00]	[-9.26, 26.54]	[3E-133, 0.77]
		mean	7.1E-05	3.3E-05	3.9128	0.1440
x_{i6}	Turnover ratio	range	[-2E-07, 8E-07]	[6E-10, 4E-08]	[-23.0, 263.8]	[0 , 0.01184]
		mean	1.7E-07	6.2E-09	45.76073	0.0002
x_{i7}	Price change	range	[-38764,62354]	[0 , 32251]	[-36.25, 10.09]	[4E-213, 0.89]
		mean	519.76	3664.40	-0.5498	0.2564
x_{i8}	Spread prior to trade	range	[-1.5677, 4.5827]	[0 , 1.7189]	[-18.89, 35.57]	[4E-207, 0.92]
		mean	0.0706	0.1231	-0.3805	0.2517
x_{i9}	Spread thereafter trade	range	[-0.0367, 0.5368]	[0.0003, 0.0653]	[-5.97, 20.72]	[4E-85, 0.996]
		mean	0.0262	0.0095	2.4358	0.2077
x_{i10}	Mid quote price prior to trade	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i11}	Mid quote price subsequent to trade	range	[-6.8458, 9.0515]	[0.0005, 3.4355]	[-18.27, 35.76]	[8E-209, 0.95]
		mean	0.0198	0.5305	0.0759	0.2324
x_{i12}	Moving average value traded	range	[-3.7E-05, 0.00]	[6.8E-08, 9.8E-05]	[-5.91, 29.26]	[1E-155, 0.99]
		mean	1.3E-05	6.5E-06	1.2271	0.3484
x_{i13}	Volatility	range	[-300.7, 603.9]	[22.4769, 592.8]	[-4.42, 6.99]	[3.9E-12, 0.95]
		mean	56.0794	105.1373	0.5052	0.3016
x_{i14}	Momentum	range	[-148.677, 317.055]	[4.2484, 254.5]	[-4.4157, 6.83]	[1.2E-11, 0.99]
		mean	1.8005	30.3232	-0.2293	0.3023
x_{i15}	Order sign	range	[-0.6483, 0.7935]	[0.0172, 0.2491]	[-18.46, 14.51]	[2.0E-72, 0.89]
		mean	-0.0138	0.0585	-0.9273	0.1926
x_{i16}	Signed volume	range	[-0.9992, 0.9742]	[0.001, 0.05]	[-817.2, 132.8]	[0 , 0.7040]
		mean	0.0165	0.0180	-11.0807	0.0614
x_{i17}	Order imbalance	range	[-0.0069, 0.9744]	[5.3E-05, 0.0534]	[-5.35, 132.82]	[5.4E-10, 0.70]
		mean	-0.0005	0.0012	-0.3316	60.2522
x_{i18}	Liquidity	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i19}	Execution time	range	[-1.1814, 1.0729]	[0.0063, 1.8805]	[-3.8784, 7.87]	[63E-15, 0.99]
		mean	0.0017	0.1037	0.2718	0.5015
x_{i20}	PIN	range	[-254.8, 137.7]	[0 , 363.2]	[-0.97, 2.08]	[0.0378, 0.96]
		mean	0.4239	23.0793	0.1424	0.6601
x_{i21}	KL divergence counts	range	[-1.7634, 2.9993]	[0.3574, 3.6915]	[-1.4159, 3.27]	[0.0014, 0.99]
		mean	0.1674	1.1565	0.17662	0.5633
x_{i22}	KL divergence volume	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
range	[0.2524, 0.9987]	[0.2431, 0.9987]	[0.14053, 17.3241]	[0.3749, 4.1622]
mean	0.3716	0.3624	0.6572	0.8107

Table 2. Johannesburg Stock Exchange (JSE). GLM parameters of large capitalization CPI stock and small capitalization ADH stock.

Feature	Feature description	Coefficient	Capitalization	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	α_0	large	-77.3893	38.9286	-1.9880	0.0471
			small	64.9161	19.1075	3.3974	0.0007
x_{i1}	Price	α_1	large	-3.4E-05	0.0005	-0.0732	0.9417
			small	-0.0758	0.0222	-3.4165	0.0006
x_{i2}	Bid price	α_2	large	0	0	-	-
			small	0.4732	0.3313	1.4281	0.1534
x_{i3}	Bid size	α_3	large	0.0004	0.0002	1.5866	0.1130
			small	4.8E-05	1.2E-05	4.1349	3.7E-05
x_{i4}	Ask price	α_4	large	-0.0005	0.0006	-0.7579	0.4487
			small	0.3759	0.3369	1.1158	0.2646
x_{i5}	Ask size	α_5	large	0.0005	0.0002	2.3523	0.0189
			small	1.6E-05	1.2E-05	1.2818	0.2000
x_{i6}	Turnover ratio	α_6	large	7.7E-08	6.0E-09	12.8032	1.6E-34
			small	8.0E-08	1.1E-09	74.3792	0
x_{i7}	Price change	α_7	large	0	0	-	-
			small	1159.0493	997.1859	1.1623	0.2452
x_{i8}	Spread prior to trade	α_8	large	9.1E-05	0.0004	0.2075	0.8357
			small	0	0	-	-
x_{i9}	Spread subsequent to trade	α_9	large	-0.0003	0.0003	-0.9090	0.3636
			small	0.0669	0.0102	6.5724	6.0E-11
x_{i10}	Mid quote price prior to trade	α_{10}	large	0	0	-	-
			small	0	0	-	-
x_{i11}	Mid quote price subsequent to trade	α_{11}	large	0.0006	0.0005	1.1258	0.2606
			small	-0.7757	0.6718	-1.1548	0.2483
x_{i12}	Moving average value traded	α_{12}	large	3.6E-07	2.3E-07	1.6016	0.1096
			small	-7.7E-07	7.1E-06	-0.1089	0.9133
x_{i13}	Volatility	α_{13}	large	480.7735	218.5557	2.1998	0.0281
			small	82.0011	130.1168	0.6302	0.5286
x_{i14}	Momentum	α_{14}	large	66.8747	37.9810	1.7607	0.0786
			small	-59.3820	18.7701	-3.1636	0.0016
x_{i15}	Order sign	α_{15}	large	-0.1172	0.0435	-2.6952	0.0072
			small	-0.0887	0.0754	-1.1761	0.2397
x_{i16}	Signed volume	α_{16}	large	0.2910	0.0299	9.7457	2.2E-21
			small	0.1651	0.0106	15.5134	8.1E-52
x_{i17}	Order imbalance	α_{17}	large	0.0031	0.0026	1.2315	0.2185
			small	0.0015	0.0009	1.7245	0.0848
x_{i18}	Liquidity	α_{18}	large	0	0	-	-
			small	0	0	-	-
x_{i19}	Execution time	α_{19}	large	0.0580	0.3013	0.1924	0.8475
			small	-0.0163	0.0270	-0.6042	0.5458
x_{i20}	PIN	α_{20}	large	0	0	-	-
			small	-21.0066	64.8719	-0.3238	0.7461
x_{i21}	KL divergence counts	α_{21}	large	-0.4113	0.4101	-1.0028	0.3162
			small	-1.0880	2.4329	-0.4472	0.6548
x_{i22}	KL divergence volume	α_{22}	large	0	0	-	-
			small	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
large	0.2828	0.2698	0.2914	0.5399
small	0.7570	0.7552	8.4242	2.9024

Table 3. Bolsa de Valores de Sao Paulo (BOVESPA). GLM parameters of the range and the average across top 42 highly liquid stocks.

Feature	Feature description	Metric	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	range	[0, 1.8862]	[0, 0.5539]	[3.41, 3.41]	[0.0007, 0.0007]
		mean	0.0449	0.0132	3.4055	0.0007
x_{i1}	Price	range	[-0.0001 , 4.1E-05]	[0 , 4E-05]	[-6.04, 1.02]	[2 E-09, 0.3065]
		mean	-2.5E-06	1.5E-06	-2.5097	0.1533
x_{i2}	Bid price	range	[0, 92.211]	[0, 125.01]	[0.13, 0.74]	[0.4608, 0.8979]
		mean	2.3379	4.0860	0.4330	0.6793
x_{i3}	Bid size	range	[0, 4.4E-05]	[0, 2.1E-06]	[20.89, 21.34]	[2.7E-98, 2E-95]
		mean	1.28E-06	6.0E-08	21.1303	10E-96
x_{i4}	Ask price	range	[-92.152 , 0]	[0 , 125.07]	[-0.74, -0.08]	[0.4612, 0.9383]
		mean	-2.2838	4.1370	-0.4071	0.6998
x_{i5}	Ask size	range	[-4.2E-06 , 0]	[0, 3.2E-06]	[-3.25, -1.29]	[0.0011, 0.1958]
		mean	-1.2E-07	8.3E-08	-2.2734	0.0985
x_{i6}	Turnover ratio	range	[0, 1.4E-05]	[0, 2.7E-05]	[0.02, 8.13]	[5 E-16, 0.9868]
		mean	3.5E-07	6.8E-07	4.0728	0.4934
x_{i7}	Price change	range	[-117416 ,163.19]	[0 , 23851]	[-4.92, 7.95]	[2.2E-15, 9E-07]
		mean	-2791	568.37	1.5119	4.3E-07
x_{i8}	Spread prior to trade	range	[-13.359 , 100.27]	[0 , 124.95]	[-0.28, 0.80]	[0.4223, 0.7789]
		mean	2.0694	4.1081	0.2609	0.6006
x_{i9}	Spread thereafter trade	range	[0, 46.708]	[0, 5.3303]	[0.10, 8.76]	[2 E-18, 0.9172]
		mean	1.1236	0.2380	4.4333	0.4586
x_{i10}	Mid quote price prior to trade	range	[-9748.5 , 160.55]	[0, 1978]	[-4.93, 10.43]	[2.6E-25, 8E-07]
		mean	-228.28	47.4611	2.7530	4.2E-07
x_{i11}	Mid quote price subsequent to trade	range	[-163.53 , 9748.6]	[0, 1978]	[-10.91, 4.93]	[1.7E-27, 8E-07]
		mean	228.21	47.4516	-2.9919	4.2E-07
x_{i12}	Moving average value traded	range	[0, 2.4E-05]	[0, 4.1E-06]	[1.55, 5.73]	[1 E-08, 0.1215]
		mean	7.0E-07	1.9E-07	3.6382	0.0607
x_{i13}	Volatility	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i14}	Momentum	range	[0, 0.0028]	[0, 0.0039]	[0.00, 0.71]	[0.4748, 0.9967]
		mean	6.7E-05	0.0001	0.3594	0.7357
x_{i15}	Order sign	range	[-0.8331 , 0]	[0, 0.1300]	[-41.09, -6.41]	[0, 1.6E-10]
		mean	-0.0355	0.0035	-23.7467	7.9E-11
x_{i16}	Signed volume	range	[0, 0.7570]	[0, 0.0111]	[4.23, 143.58]	[0, 2.4E-05]
		mean	0.0191	0.0004	73.9045	1.2E-05
x_{i17}	Order imbalance	range	[-0.0001 , 0]	[0, 2.7E-05]	[-4.33, -0.89]	[2 E-05, 0.3712]
		mean	-2.8E-06	7.3E-07	-2.6140	0.1856
x_{i18}	Liquidity	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i19}	Execution time	range	[0, 2.6615]	[0, 0.9772]	[0.15, 2.72]	[0.0065, 0.8845]
		mean	0.0636	0.0247	1.4345	0.4455
x_{i20}	PIN	range	[0, 0]	[0 , 0]	-	-
		mean	0	0	-	-
x_{i21}	KL divergence counts	range	[-0.9385 , 0]	[0, 5.7105]	[-0.16, -0.16]	[0.8695, 0.8695]
		mean	-0.0223	0.1360	-0.1644	0.8695
x_{i22}	KL divergence volume	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
range	[0.1851 , 0.6575]	[0.1832 , 0.6571]	[0.5441 , 10420]	[0.7376, 102.08]
mean	0.1851	0.1832	13.2603	3.6414

Table 4.  Bolsa de Valores de Sao Paulo (BOVESPA). GLM parameters of large capitalization VVA stock and small capitalization CGA stock.

Feature	Feature description	Coefficient	Capitalization	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	α_0	large	1.8862	0.5539	3.4055	0.0007
			small	0	0	-	-
x_{i1}	Price	α_1	large	4.1E-05	4 E-05	1.0226	0.3065
			small	-0.0001	2.4E-05	-6.0420	1.6E-09
x_{i2}	Bid price	α_2	large	5.9820	46.5968	0.1284	0.8979
			small	92.2113	125.0141	0.7376	0.4608
x_{i3}	Bid size	α_3	large	4.4E-05	2.1E-06	21.3758	2.7E-98
			small	9.5E-06	4.5E-07	20.8848	2 E-95
x_{i4}	Ask price	α_4	large	-3.7681	48.6875	-0.0774	0.9383
			small	-92.1519	125.0664	-0.7368	0.4612
x_{i5}	Ask size	α_5	large	-4.2E-06	3.2E-06	-1.2937	0.1958
			small	-7.9E-07	2.4E-07	-3.2531	0.0011
x_{i6}	Turnover ratio	α_6	large	4.5E-07	2.7E-05	0.0166	0.9868
			small	1.4E-05	1.8E-06	8.1289	4.7E-16
x_{i7}	Price change	α_7	large	163.1925	20.5358	7.9467	2.2E-15
			small	-117416	23850	-4.9229	8.6E-07
x_{i8}	Spread prior to trade	α_8	large	-13.3591	47.5898	-0.2807	0.77894
			small	100.2721	124.9507	0.8025	0.4223
x_{i9}	Spread subsequent to trade	α_9	large	46.7085	5.3303	8.7627	2.4E-18
			small	0.4847	4.6637	0.1039	0.9172
x_{i10}	Mid quote price prior to trade	α_{10}	large	160.5550	15.3873	10.4343	2.6E-25
			small	-9748.5038	1977.9779	-4.9285	8.4E-07
x_{i11}	Mid quote price subsequent to trade	α_{11}	large	-163.5307	14.9858	-10.9124	1.7E-27
			small	9748.5675	1977.9805	4.9285	8.4E-07
x_{i12}	Moving average value traded	α_{12}	large	2.4E-05	4.1E-06	5.7275	1.1E-08
			small	6 E-06	3.8E-06	1.5488	0.1215
x_{i13}	Volatility	α_{13}	large	0	0	-	-
			small	0	0	-	-
x_{i14}	Momentum	α_{14}	large	0.0028	0.0039	0.7147	0.4748
			small	7.7E-06	0.0018	0.0042	0.9967
x_{i15}	Order sign	α_{15}	large	-0.8331	0.1300	-6.4077	1.6E-10
			small	-0.6589	0.0160	-41.0858	0
x_{i16}	Signed volume	α_{16}	large	0.0470	0.0111	4.2283	2.4E-05
			small	0.7570	0.0053	143.5807	0
x_{i17}	Order imbalance	α_{17}	large	-0.0001	2.7E-05	-4.3338	1.5E-05
			small	-3.6E-06	4.1E-06	-0.8943	0.3712
x_{i18}	Liquidity	α_{18}	large	0	0	-	-
			small	0	0	-	-
x_{i19}	Execution time	α_{19}	large	0.0090	0.0618	0.1453	0.8845
			small	2.6615	0.9772	2.7237	0.0065
x_{i20}	PIN	α_{20}	large	0	0	-	-
			small	0	0	-	-
x_{i21}	KL divergence counts	α_{21}	large	0	0	-	-
			small	-0.9385	5.7105	-0.1644	0.8695
x_{i22}	KL divergence volume	α_{22}	large	0	0	-	-
			small	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
large	0.1851	0.1832	13.2603	3.6415
small	0.6575	0.6571	0.7792	0.8827

Table 5. Moscow Exchange (MOEX). GLM parameters of the range and the average across top 30 highly liquid stocks.

Feature	Feature description	Metric	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	range	[-24480.8, 3850.5]	[0, 14324.7]	[-2.6652, 21.881]	[2E-104, 1]
		mean	-687.42	1636.8	0.8352	0.8967
x_{i1}	Price	range	[-2.2E-05 , 1.1E-05]	[1.2E-09 , 0.0016]	-[-0.5140, 0.5788]	[0.5627, 1]
		mean	-1.2E-07	7.2E-05	0.0135	0.9624
x_{i2}	Bid price	range	[-85.795, 38.098]	[0, 32.249]	[-2.6604, 4.9981]	[5.9E-07 , 1]
		mean	-1.5938	4.2953	0.1457	0.8752
x_{i3}	Bid size	range	[-8.9E-17, 0.0026]	[4.3E-07, 9.6E-05]	[-6.5E-12, 277.94]	[0, 1]
		mean	0.0003	1.3E-05	19.083	0.8667
x_{i4}	Ask price	range	[-87.870, 37.511]	[0, 86.882]	[-2.6695, 5.0029]	[5.8E-07, 1]
		mean	-1.9307	13.667	0.0709	0.9237
x_{i5}	Ask size	range	[-7.6E-06, 7.4E-05]	[3.2E-07 , 6.7E-05]	[-0.3293, 10.827]	[4.2E-27, 1]
		mean	2.2E-06	1.1E-05	0.3494	0.9576
x_{i6}	Turnover ratio	range	[-2.1E-07, 4.6E-07]	[1.3E-09, 5.5E-07]	[-1.1493, 1.0689]	[0.2505, 1]
		mean	1.4E-08	6.0E-08	0.0265	0.9302
x_{i7}	Price change	range	[-24155, 3807.1]	[0 , 14177]	[-2.6628, 5.0403]	[4.8E-07, 1]
		mean	-744.14	1624.9	0.0723	0.9236
x_{i8}	Spread prior to trade	range	[-0.0193, 3.5258]	[0 , 44.124]	[-0.0011, 0.2564]	[0.7971, 1]
		mean	0.1169	4.6909	0.0196	0.9844
x_{i9}	Spread thereafter trade	range	[-0.1561, 0.8007]	[0, 12.496]	[-0.8688, 1.8250]	[0.0680, 1]
		mean	0.0285	0.5417	0.0947	0.9145
x_{i10}	Mid quote price prior to trade	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i11}	Mid quote price subsequent to trade	range	[-75.5934, 173.67]	[0, 86.888]	[-5.0009, 2.6653]	[5.8E-07, 1]
		mean	3.5254	17.111	-0.0736	0.9210
x_{i12}	Moving average value traded	range	[-6.1E-05, 3.3E-07]	[2E-11, 5.9E-05]	[-16.703, 0.2672]	[1.1E-61, 1]
		mean	-2.0E-06	3.9E-06	-0.5532	0.9523
x_{i13}	Volatility	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i14}	Momentum	range	[-7.5E-06, 0.0019]	[1.1E-05, 0.0033]	[-0.0070, 1.0908]	[0.2754, 1]
		mean	0.0002	0.0005	0.0770	0.9450
x_{i15}	Order sign	range	[-5.0920, 1.8976]	[0.0012, 0.0265]	[-234.67, 410.14]	[0, 1]
		mean	-0.0632	0.0076	7.5644	0.6477
x_{i16}	Signed volume	range	[-1, 1]	[5.4E-05, 0.0021]	[-18341, 18501]	[0, 0]
		mean	-0.0147	0.0007	-199.57	0
x_{i17}	Order imbalance	range	[-0.0001, 6.3E-09]	[3.6E-08, 4.7E-05]	[-5.2130, 0.0092]	[1.9E-07, 1]
		mean	-5.2E-06	5.8E-06	-0.2075	0.9408
x_{i18}	Liquidity	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i19}	Execution time	range	[-0.0778, 1.4E-05]	[5.7E-05, 2.27035]	[-0.1071, 0.0024]	[0.9147, 1]
		mean	-0.0026	0.1487	-0.0046	0.9962
x_{i20}	PIN	range	[0, 0]	[0 , 0]	-	-
		mean	0	0	-	-
x_{i21}	KL divergence counts	range	[-3851.0, 24479.5]	[0, 14325]	[-5.0018, 2.6651]	[5.8E-07, 1]
		mean	687.69	1636.8	-0.0758	0.9275
x_{i22}	KL divergence volume	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
range	[0.8804, 1]	[0.8800, 1]	[1.9E-05, 0.6990]	[0.0043, 0.8361]
mean	0.9999	0.9999	0.0008	0.0291

Table 6. Moscow Exchange (MOEX). GLM parameters of large capitalization VTB stock and small capitalization TRN stock.

Feature	Feature description	Coefficient	Capitalization	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	α_0	large	5.0920	0.2327	21.8807	2E-104
			small	0	78.0179	0	1
x_{i1}	Price	α_1	large	1.9E-22	1.2E-09	1.6E-13	1
			small	1.0E-16	0.0016	6.4E-14	1
x_{i2}	Bid price	α_2	large	0	0	-	-
			small	0	0	-	-
x_{i3}	Bid size	α_3	large	9.6E-19	4.3E-07	2.2E-12	1
			small	-8.9E-17	9.6E-05	-9.3E-13	1
x_{i4}	Ask price	α_4	large	0	0	-	-
			small	-1.2E-16	0.0005	-2.5E-13	1
x_{i5}	Ask size	α_5	large	1.3E-19	1.1E-06	1.2E-13	1
			small	7.4E-18	6.7E-05	1.1E-13	1
x_{i6}	Turnover ratio	α_6	large	4.9E-22	2.4E-08	1.9E-14	1
			small	6.7E-23	5.7E-09	1.2E-14	1
x_{i7}	Price change	α_7	large	0	0	-	-
			small	-1.9E-11	74.4	-2.5E-13	1
x_{i8}	Spread prior to trade	α_8	large	0	0	-	-
			small	6.1E-17	0.0002	2.6E-13	1
x_{i9}	Spread subsequent to trade	α_9	large	0	0	-	-
			small	1.3E-17	4.4E-06	2.9E-12	1
x_{i10}	Mid quote price prior to trade	α_{10}	large	0	0	-	-
			small	0	0	-	-
x_{i11}	Mid quote price subsequent to trade	α_{11}	large	0	0	-	-
			small	1.2E-16	0.0005	2.5E-13	1
x_{i12}	Moving average value traded	α_{12}	large	-2E-23	2E-11	-1.0E-12	1
			small	3.5E-18	5.7E-05	5.9E-14	1
x_{i13}	Volatility	α_{13}	large	0	0	-	-
			small	0	0	-	-
x_{i14}	Momentum	α_{14}	large	-8.7E-18	3E-05	-3E-13	1
			small	-1.1E-16	0.0015	-7.0E-14	1
x_{i15}	Order sign	α_{15}	large	-5.0920	0.0217	-234.67	0
			small	1.9E-11	0.0067	2.9E-09	1
x_{i16}	Signed volume	α_{16}	large	1	0.0005	1992.9	0
			small	-1	0.0002	-5039.9	0
x_{i17}	Order imbalance	α_{17}	large	-4.4E-20	5.8E-07	-7.6E-14	1
			small	-5.8E-20	7.0E-07	-8.2E-14	1
x_{i18}	Liquidity	α_{18}	large	0	0	-	-
			small	0	0	-	-
x_{i19}	Execution time	α_{19}	large	1.7E-17	0.0058	2.9E-15	1
			small	-3.2E-19	0.0008	-4E-16	1
x_{i20}	PIN	α_{20}	large	0	0	-	-
			small	0	0	-	-
x_{i21}	KL divergence counts	α_{21}	large	-5E-14	0.2338	-2.1E-13	1
			small	1.9E-11	78.0200	2.5E-13	1
x_{i22}	KL divergence volume	α_{22}	large	0	0	-	-
			small	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
large	0.9981	0.9981	0.1083	0.3291
small	0.9999	0.9997	0.0017	0.0409

Table 7. National Stock Exchange of India (NSE). GLM parameters of the range and the average across top 43 highly liquid stocks.

Feature	Feature description	Metric	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	range	[-48244.05, 301019]	[0, 158781]	[-7.61, 4.69]	[3.0E-14, 0.9984]
		mean	4826.4	0	-	-
x_{i1}	Price	range	[-0.1685 , 0.0211]	[1.2E-05 , 0.04]	[-24.7, 21.1]	[4E-132 , 0.9757]
		mean	-0.0041	0	-	-
x_{i2}	Bid price	range	[-68.887, 110.34]	[0, 99.5]	[-7.95, 2.91]	[2.0E-15 , 0.9770]
		mean	1.5663	0	-	-
x_{i3}	Bid size	range	[4.2E-05, 0.0124]	[4.3E-07, 0.00]	[1.45, 171]	[0 , 0.1468]
		mean	0.0018	0	-	-
x_{i4}	Ask price	range	[-1053.2, 109.33]	[0 , 3005.2]	[-7.64, 7.97]	[1.6E-15, 0.9885]
		mean	0	0	-	-
x_{i5}	Ask size	range	[-0.0004, 0.0008]	[3.6E-07, 0.00]	[-6.21, 4.47]	[5.4E-10, 0.9437]
		mean	0	0	-	-
x_{i6}	Turnover ratio	range	[-7.2E-05, 0.0001]	[6E-08, 6E-05]	[-18.5, 25.6]	[5E-142, 0.9470]
		mean	0	0	-	-
x_{i7}	Price change	range	[-49424.8, 300290]	[30.9 , 159764]	[-12.2, 4.63]	[5.5E-34, 0.9990]
		mean	0	0	-	-
x_{i8}	Spread prior to trade	range	[-20.877, 522.67]	[0 , 1501.3]	[-4.70, 5.81]	[6.4E-09, 0.9914]
		mean	0	0	-	-
x_{i9}	Spread thereafter trade	range	[-1.1184, 6.8567]	[0.00, 6.65]	[-6.98, 22.1]	[7.4E-106, 0.9769]
		mean	0	0	-	-
x_{i10}	Mid quote price prior to trade	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i11}	Mid quote price subsequent to trade	range	[-219.86, 1050.6]	[0.00, 3005]	[-4.69, 7.61]	[3.0E-14, 0.9951]
		mean	0	0	-	-
x_{i12}	Moving average value traded	range	[-0.0002, 0.0007]	[7.0E-07, 0.00]	[-9.17, 35.8]	[3.9E-266, 0.9744]
		mean	0	0	-	-
x_{i13}	Volatility	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i14}	Momentum	range	[-0.0049, 0.0031]	[5.1E-05, 0.00]	[-9.83, 3.52]	[10E-23, 0.9809]
		mean	0	0	-	-
x_{i15}	Order sign	range	[-1.5531, 1.5733]	[0.01, 0.50]	[-23.1, 62.0]	[0, 0.8375]
		mean	0	0	-	-
x_{i16}	Signed volume	range	[-0.9964, 0.5008]	[0.00, 0.05]	[-718, 70.1]	[0, 0.5586]
		mean	0	0	-	-
x_{i17}	Order imbalance	range	[-0.0070, 0.0003]	[1.4E-06, 0.00]	[-13.0, 29.5]	[1.5E-184, 0.9863]
		mean	0	0	-	-
x_{i18}	Liquidity	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i19}	Execution time	range	[-0.0047, 0.0139]	[0.00, 0.03]	[-2.95, 1.04]	[0.0032, 0.9971]
		mean	0	0	-	-
x_{i20}	PIN	range	[0, 0]	[0 , 0]	-	-
		mean	0	0	-	-
x_{i21}	KL divergence counts	range	[-300746, 48369]	[0, 158790]	[-4.69, 7.61]	[3.0E-14, 0.9905]
		mean	0	0	-	-
x_{i22}	KL divergence volume	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
range	[0.3252, 0.9962]	[0.2939, 0.9962]	[0.0164, 21.505]	[0.1282, 4.6373]
mean	0.8778	0.8777	0.9287	0.9637

Table 8. National Stock Exchange of India (NSE). GLM parameter fitting of large capitalization STC stock and small capitalization VJB stock.

Feature	Feature description	Coefficient	Capitalization	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	α_0	large	1150.3	495.2	2.3227	0.0202
			small	-48244.0	138019.8	-0.3495	0.7271
x_{i1}	Price	α_1	large	-0.0004	0.0003	-1.3296	0.1837
			small	0.0017	0.0028	0.6162	0.5385
x_{i2}	Bid price	α_2	large	4.8855	1.9380	2.5209	0.0117
			small	0	0	-	-
x_{i3}	Bid size	α_3	large	0.0026	3.1E-05	84.025	0
			small	6E-05	5.8E-06	10.264	1.2E-19
x_{i4}	Ask price	α_4	large	4.0976	1.9356	2.1170	0.0343
			small	-1053.2	3005.2	-0.3505	0.7264
x_{i5}	Ask size	α_5	large	4.6E-05	3.4E-05	1.3723	0.1700
			small	-5.3E-06	1.8E-06	-3.0032	0.0031
x_{i6}	Turnover ratio	α_6	large	3.1E-06	2.3E-06	1.3706	0.1705
			small	-3.7E-05	6.0E-05	-0.6161	0.5386
x_{i7}	Price change	α_7	large	1179.4	516.9	2.2817	0.0225
			small	-49424.8	138182.1	-0.3577	0.7210
x_{i8}	Spread prior to trade	α_8	large	0	0	-	-
			small	522.7	1501.3	0.3482	0.7281
x_{i9}	Spread subsequent to trade	α_9	large	-0.0140	0.1521	-0.0923	0.9265
			small	6.4158	6.6500	0.9648	0.3360
x_{i10}	Mid quote price prior to trade	α_{10}	large	0	0	-	-
			small	0	0	-	-
x_{i11}	Mid quote price subsequent to trade	α_{11}	large	-8.9790	3.8668	-2.3221	0.0202
			small	1050.6	3005.3	0.3496	0.7271
x_{i12}	Moving average value traded	α_{12}	large	-6.3E-05	6.8E-06	-9.1716	5.4E-20
			small	-1.5E-06	1.3E-06	-1.2047	0.2300
x_{i13}	Volatility	α_{13}	large	0	0	-	-
			small	0	0	-	-
x_{i14}	Momentum	α_{14}	large	-0.0014	0.0011	-1.3439	0.1790
			small	-0.0002	0.0021	-0.1194	0.9051
x_{i15}	Order sign	α_{15}	large	-0.0227	0.0204	-1.1148	0.2650
			small	-1.5531	0.1721	-9.0242	3.2E-16
x_{i16}	Signed volume	α_{16}	large	0.0467	0.0073	6.3957	1.7E-10
			small	0.4792	0.0460	10.409	5E-20
x_{i17}	Order imbalance	α_{17}	large	0.0001	2E-05	6.7506	1.54E-11
			small	0.0003	0.0007	0.4067	0.6847
x_{i18}	Liquidity	α_{18}	large	0	0	-	-
			small	0	0	-	-
x_{i19}	Execution time	α_{19}	large	-0.0006	0.0012	-0.5286	0.5971
			small	-0.0019	0.0089	-0.2179	0.8278
x_{i20}	PIN	α_{20}	large	0	0	-	-
			small	0	0	-	-
x_{i21}	KL divergence counts	α_{21}	large	-1150.1	495.2	-2.3223	0.0202
			small	48369.1	138012.4	0.3505	0.7264
x_{i22}	KL divergence volume	α_{22}	large	0	0	-	-
			small	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
large	0.3829	0.3821	3.0663	1.7511
small	0.8337	0.8190	0.5250	0.7246

Table 9. Shanghai Stock Exchange (SSE). . GLM parameters of the range and the average across top 54 highly liquid stocks.

Feature	Feature description	Metric	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	range	[-2091.2, 214.75]	[0, 4599.7]	[-2.3138, 9.8733]	[7E-23, 0.9922]
		mean	-77.0779	316.50	1.1410	0.4071
x_{i1}	Price	range	[-0.0004 , 8.0E-05]	[0 , 6.3E-05]	[-11.536, 13.117]	[7E-39, 0.9596]
		mean	-3.1E-05	1.3E-05	-1.5819	0.2343
x_{i2}	Bid price	range	[-85.411, 78.223]	[0, 227.32]	[-0.6268, 6.6686]	[3E-11, 0.9870]
		mean	0.0872	13.179	0.8238	0.5423
x_{i3}	Bid size	range	[0, 8.4E-05]	[0, 2.7E-06]	[9.8747, 34.565]	[5E-250, 1E-22]
		mean	1.6E-05	7.0E-07	22.281	2.0E-24
x_{i4}	Ask price	range	[-92.050, 85.163]	[0 , 344.7031]	[-6.6529, 1.8235]	[3.0E-11, 0.9920]
		mean	-6.0947	43.964	-0.5264	0.5643
x_{i5}	Ask size	range	[-5.0E-06, 1.1E-05]	[0, 2.7E-06]	[-3.2091, 6.3616]	[2.1E-10, 0.9358]
		mean	9.9E-07	6.8E-07	0.8051	0.2400
x_{i6}	Turnover ratio	range	[-3.9E-06, 4.5E-05]	[0, 4.1E-06]	[-9.3223, 14.091]	[1.2E-44, 0.8451]
		mean	2.5E-06	1.1E-06	1.6776	0.2358
x_{i7}	Price change	range	[-119523,339.42]	[0 , 29680]	[-25.616, 7.2690]	[8E-140, 0.9466]
		mean	-2937.6600	732.69	-3.4739	0.1745
x_{i8}	Spread prior to trade	range	[-110.56, 77.203]	[0 , 227.05]	[-7.7179, 1.5072]	[1.3E-14, 0.9629]
		mean	-17.1298	33.719	-1.5873	0.3074
x_{i9}	Spread thereafter trade	range	[-2.4998, 187.13]	[0, 30.984]	[-1.9522, 21.952]	[4E-104, 0.6082]
		mean	29.6358	5.2726	5.5162	0.0468
x_{i10}	Mid quote price prior to trade	range	[-4447.1, 110.95]	[0, 1117.0]	[-23.624, 8.6036]	[8E-120, 0.9581]
		mean	-146.8893	0	-1.5842	0.2934
x_{i11}	Mid quote price subsequent to trade	range	[-111.05, 4447.1]	[0, 1117.0]	[-8.5679, 23.642]	[5E-120, 0.9863]
		mean	152.5327	47.4082	2.0453	0.1896
x_{i12}	Moving average value traded	range	[-8.0E-06, 3.6E-05]	[0, 1.2E-05]	[-11.494, 57.002]	[0 , 0.9631]
		mean	1.4E-06	9.0E-07	1.8012	0.3233
x_{i13}	Volatility	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i14}	Momentum	range	[-0.0043, 0.0242]	[0, 0.0093]	[-2.6381, 14.992]	[2.4E-50, 0.9918]
		mean	0.0024	0.0018	1.2927	0.3784
x_{i15}	Order sign	range	[-0.8731, 0.7090]	[0, 0.0486]	[-70.863, 14.584]	[0 , 0.0163]
		mean	-0.5574	0.0209	-31.933	0.0003
x_{i16}	Signed volume	range	[-0.7426, 0.9993]	[0, 0.0106]	[-109.94, 4222.5]	[0 , 0.3855]
		mean	0.7900	0.0045	327.95	0.0074
x_{i17}	Order imbalance	range	[-0.0006, 0.0002]	[0, 0.0004]	[-6.6841, 9.5491]	[2E-21, 0.9875]
		mean	-4.4E-06	2.7E-05	0.5959	0.2705
x_{i18}	Liquidity	range	[0, 0]	[0, 0]	-	-
		mean	0	0	-	-
x_{i19}	Execution time	range	[-0.0064, 0.0104]	[0, 0.0172]	[-2.4533, 2.0931]	[0.01417, 0.9974]
		mean	0.0001	0.0047	-0.0524	0.7259
x_{i20}	PIN	range	[0, 0]	[0 , 0]	-	-
		mean	0	0	-	-
x_{i21}	KL divergence counts	range	[-9.9781, 144.34]	[0, 431.1649]	[-2.0845, 5.1614]	[2.5E-07, 0.9606]
		mean	6.5462	9.3598	1.1263	0.3010
x_{i22}	KL divergence volume	range	[-220.43, 2093.7]	[0, 4599.7205]	[-0.3160, 1.5758]	[0.1151, 0.9921]
		mean	75.9752	307.36	0.4225	0.6721

	Ordinary R^2	Adjusted R^2	MSE	RMSE
range	[0.1876, 0.9998]	[0.1860, 0.9998]	[0.0946, 15.7656]	[0.3076, 3.9706]
mean	0.6529	0.6521	4.3651	2.0893

Table 10. Shanghai Stock Exchange (SSE). GLM parameter fitting of large capitalization 600803.SS and small capitalization 900948.SS stock.

Feature	Feature description	Coefficient	Capitalization	Coefficient value(s)	SE	tStat	pValue
x_{i0}	Bias	α_0	large	-0.7074	2.3795	-0.2973	0.7662
			small	-4.0795	2.8350	-1.4390	0.1502
x_{i1}	Price	α_1	large	2.4E-06	4.4E-06	0.5595	0.5758
			small	3.4E-06	2.3E-05	0.1474	0.8828
x_{i2}	Bid price	α_2	large	0	0	-	-
			small	0	0	-	-
x_{i3}	Bid size	α_3	large	1.3E-06	5.7E-08	22.789	3E-111
			small	3.9E-05	1.7E-06	22.583	3E-109
x_{i4}	Ask price	α_4	large	-0.0400	0.2500	-0.1600	0.8729
			small	0.0012	0.0425	0.0287	0.9771
x_{i5}	Ask size	α_5	large	-2.4E-08	5.4E-08	-0.4503	0.6525
			small	2.5E-06	1.9E-06	1.3033	0.1925
x_{i6}	Turnover ratio	α_6	large	6.6E-07	1.1E-06	0.6295	0.5290
			small	-1.2E-06	6.5E-07	-1.8484	0.0646
x_{i7}	Price change	α_7	large	-613.54	309.86	-1.9801	0.0477
			small	-465.13	305.19	-1.5240	0.1275
x_{i8}	Spread prior to trade	α_8	large	-47.545	18.500	-2.5700	0.0102
			small	2.1184	2.1198	0.9993	0.3177
x_{i9}	Spread subsequent to trade	α_9	large	55.974	15.990	3.5007	0.0005
			small	2.4943	2.0387	1.2235	0.2212
x_{i10}	Mid quote price prior to trade	α_{10}	large	-107.41	82.193	-1.3069	0.1913
			small	-20.741	13.497	-1.5367	0.1244
x_{i11}	Mid quote price subsequent to trade	α_{11}	large	107.71	82.170	1.3108	0.1900
			small	20.923	13.496	1.5504	0.1211
x_{i12}	Moving average value traded	α_{12}	large	2E-08	3.5E-08	0.5632	0.5733
			small	-1.5E-07	4.5E-07	-0.3345	0.7380
x_{i13}	Volatility	α_{13}	large	0	0	-	-
			small	0	0	-	-
x_{i14}	Momentum	α_{14}	large	1.1E-05	0.0007	0.0165	0.9868
			small	0.0071	0.0027	2.6120	0.0090
x_{i15}	Order sign	α_{15}	large	-0.6374	0.0178	-35.8540	2E-260
			small	-0.6156	0.0203	-30.306	5E-190
x_{i16}	Signed volume	α_{16}	large	0.8971	0.0045	197.5142	0
			small	0.9039	0.0045	201.64	0
x_{i17}	Order imbalance	α_{17}	large	-1.8E-05	1.2E-05	-1.5445	0.1225
			small	-4.2E-07	2.1E-05	-0.0201	0.9839
x_{i18}	Liquidity	α_{18}	large	0	0	-	-
			small	0	0	-	-
x_{i19}	Execution time	α_{19}	large	0.0021	0.0063	0.3377	0.7356
			small	0.0009	0.0039	0.2196	0.8262
x_{i20}	PIN	α_{20}	large	0	0	-	-
			small	0	0	-	-
x_{i21}	KL divergence counts	α_{21}	large	0	0	-	-
			small	0	0	-	-
x_{i22}	KL divergence volume	α_{22}	large	0	0	-	-
			small	0	0	-	-

	Ordinary R^2	Adjusted R^2	MSE	RMSE
large	0.8748	0.8745	1.7817	1.3348
small	0.8668	0.8665	2.1384	1.4623

Table 11. Johannesburg Stock Exchange (JSE). ANN parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Best Epoch	[1 , 1000]	154.7619	17	1000
Best Performance	[3.7527E-09, 623.70]	29.16	0.5381	4.21E-08
Best Validation Performance	[2.3542E-08, 2029.48]	97.44	1.4485	6.76E-08
Best Test Performance	[1.1619E-05, 2822.49]	269.38	0.5537	1.1619E-05

	R^2	MSE
small	0.9983	0.6771
large	1	1.7842E-06
range	[0.0313, 1]	[1.7842E-06, 549.93]]
mean	0.8976	75.4800

Table 12. Bolsa de Valores de Sao Paulo (BOVESPA). ANN parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Best Epoch	[0, 1000]	24.4048	25	1000
Best Performance	[3.8E-07, 180792938]	13543301	0.0913	3.8E-07
Best Validation Performance	[1.3E-06, 178805408]	13511823	2.8174	1.3E-06
Best Test Performance	[1 E-06, 177435288]	13498304	416.3422	1.0348E-06

	R^2	MSE
small	0.9930	62.9393
large	1	6.1173E-07
range	[4.6E-06, 1]	[6.1E-07, 179991220]]
mean	0.0524	13531831.42

Table 13. Moscow Exchange (MOEX). ANN parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Best Epoch	[0 , 0]	0	0	0
Best Performance	[76.452, 2417454.8]	235978.7	4776.7	293400.4
Best Validation Performance	[67.706, 2351074.8]	233902.0	4739.9	293795.7
Best Test Performance	[72.073, 2421946.8]	235890.4	4786.2	291297.7

	R^2	MSE
small	0.0026	4772.6
large	0.0151	293144.3
range	[9.4E-06 , 0.0596]	[74.484 , 2408169.8]]
mean	0.0137	235653.9

Table 14. National Stock Exchange of India (NSE). ANN parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Best Epoch	[6 , 1000]	326.47	87	8
Best Performance	[1.2E-09, 1.6590]	0.1339	1.1E-05	1.5885
Best Validation Performance	[1.5E-08, 4.6438]	0.2137	2.2E-05	0.0258
Best Test Performance	[4.9E-09, 59.218]	3.2533	0.0001	0.0238

	R^2	MSE
small	1	3.0E-05
large	0.8154	1.1194
range	[0.6010 , 1]	[1.2E-08 , 8.8851]]
mean	0.9645	0.6138

Table 15. Shanghai Stock Exchange (SSE). ANN parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Best Epoch	[0, 1000]	139.5	15	23
Best Performance	[5.1E-08, 350.46]	12.523	0.4171	0.0141
Best Validation Performance	[6.8E-08, 323.35]	14.296	1.6384	3.0289
Best Test Performance	[5.9E-07, 515.01]	28.1612	0.0765	0.0121

	R^2	MSE
small	0.9545	0.5492
large	0.9761	0.4661
range	[0.0121, 1]]	[1.6E-07, 337.39]]
mean	0.8787	15.134

Table 16. Johannesburg Stock Exchange (JSE). SVM parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Gap	[0.0008, 0.0054]	0.0010	0.0010	0.0008
Delta Gradient	[0.0015, 0.1719]	0.0391	0.0264	0.0488
Largest KK Violation	[0.0010, 0.1288]	0.0292	0.0168	0.0392
Objective	[-1281.2, -5.19]	-228.8	-221.5	-290.1
Bias	[-0.320, -0.010]	-0.111	-0.0571	-0.3095
Epsilon	[0.0017, 0.0688]	0.0222	0.0091	0.0541

	R^2	MSE
small	-	0.3416
large	-	0.9213
range	-	[0.0015, 1.3593]
mean	-	0.4284

Table 17. Bolsa de Valores de Sao Paulo (BOVESPA). SVM parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Gap	[0.0006, 0.0531]	0.0047	0.0010	0.0008
Delta Gradient	[0.0012, 0.2497]	0.0489	0.1921	0.1056
Largest KK Violation	[0.0008, 0.1565]	0.0386	0.1565	0.0951
Objective	[-725.5780, -6.7402]	-210.2948	-580.5890	-291.37014
Bias	[-0.0556, 0.0032]	-0.0230	-0.0294	-0.0347
Epsilon	[0.0002, 0.0487]	0.0135	0.0228	0.0200
	R^2		MSE	
small	0.9511		633.0148	
large	0.2710		11.7123	
range	[0.0011, 0.9999]		[0.8151, 30278.64]	
mean	0.7554		2470.4343	

Table 18. Moscow Exchange (MOEX). SVM parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Gap	[0.0006, 0.0651]	0.0033	0.0009	0.0009
Delta Gradient	[0.0047, 0.0947]	0.02997	0.0065	0.0160
Largest KK Violation	[0.0033, 0.0785]	0.0242	0.0050	0.0099
Objective	[-1387.4, -3.1145]	-198.32	-5.8248	-59.226
Bias	[-0.1483, 0.0413]	-0.0218	0.0007	-0.0030
Epsilon	[0.0013, 0.0775]	0.0190	0.0139	0.0154
	R^2		MSE	
small	0.9996		0.0033	
large	0.9810		0.0443	
range	[0.0496, 0.9996]		[0.0004, 0.5566]	
mean	0.77823		0.1496	

Table 19.  National Stock Exchange of India (NSE). SVM parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Gap	[0.0005, 0.0010]	0.0009	0.0006	0.0009
Delta Gradient	[0.0139, 0.5894]	0.1200	0.0264	0.3012
Large KK Violation	[0.0110, 0.5634]	0.1048	0.0208	0.2817
Objective	[-3848.95, -7.5658]	-1075.72	-24.57	-3201.65
Bias	[-0.2536, 0.0199]	-0.0520	0.0199	-0.1097
Epsilon	[0.0068, 0.1040]	0.0297	0.1040	0.0272
	R^2		MSE	
small	0.7798		0.2754	
large	0.2584		0.7125	
range	[0.0108, 0.9936]		[0.0039, 1.9587]	
mean	0.5339		0.5156	

Table 20.  Shanghai Stock Exchange (SSE). SVM parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
Gap	[0.0008, 0.0697]	0.0022	0.0010	0.0010
Delta Gradient	[0.0045, 0.1324]	0.0556	0.0709	0.0410
Largest KK Violation	[0.0032, 0.1139]	0.0455	0.0633	0.0363
Objectives	[-1116.74, -11.896]	-489.86	-551.22	-576.34
Bias	[-0.1075, -0.0029]	-0.0451	-0.0606	-0.0421
Epsilon	[0.0003, 0.0410]	0.0150	0.0128	0.0164
	R^2		MSE	
small	0.4442		0.1582	
large	0.9198		0.1426	
range	[0.2338, 0.9982]		[0.0002, 2.6391]	
mean	0.7703		0.2404	

Table 21.  Johannesburg Stock Exchange (JSE). RF parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
ResubLoss	[0.0293, 158.5]	0.1460	7.2578	0.0645
	R^2		MSE	
small	0.5012		16.1866	
large	0.0172		1023.4	
range	[0.0007, 0.8155]		[0.3505, 2755.4]	
mean	0.0897		734.6	

Table 22.  Bolsa de Valores de Sao Paulo (BOVESPA). RF parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
ResubLoss	[0.0164, 6721.6039]	6.2228	0.3272	6.2228
	R^2		MSE	
small	0.0513		698.4299	
large	0.1956		28.1800	
range	[0.0151, 0.9690]		[0.8416, 35473.5393]	
mean	0.1956		28.1800	

Table 23.  Moscow Exchange (MOEX). RF parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
ResubLoss	[0.0191, 318.68]	0.5083	2.7065	1.3548
	R^2		MSE	
small	0.8026		2.6796	
large	0.4985		25.563	
range	[0.0753, 0.9774]		[0.1582, 188.023]	
mean	0.5336		13.446	

Table 24.  National Stock Exchange of India (NSE). RF parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
ResubLoss	[0.0092, 0.2641]	0.1746	0.1277	0.2641
	R^2		MSE	
small	0.8849		0.8223	
large	0.8172		1.1059	
range	[0.1393, 0.8172]		[0.1025, 1.1059]	
mean	0.8102		2.1144	

Table 25.  Shanghai Stock Exchange (SSE). RF parameter fitting of range, mean, small cap stock and and large cap stock.

Metric	Range	Mean	Small Capitalization Stock	Large Capitalization Stock
ResubLoss	[0.2087, 956.91]	2.0156	2.9005	3.0392
	R^2		MSE	
small	0.7748		4.2119	
large	0.6627		7.4560	
range	[0.1409, 0.9742]		[0.0944, 856.01]	
mean	0.2501		158.36	

Appendices

Appendix A. Market Features and Average Impact Curves

The nomenclature provides market features, description and reference; meanwhile, the graphs of Figure A.21 -A.25 are included as evidence to support some of the views of Section 3.

Nomenclature

Price	The matching of a limit order and a market order of opposite signs as a consequence of the interplay between the order book, and order flow. Potters *et al.* (2003)
Bid price	The price in cents at which the market maker buys a specified-security as a specified time instant. Potters *et al.* (2003)

Bid size	The number of shares in lots that are quoted at the bid price. Potters *et al.* (2003)
Ask price	The price in cents at which the market maker sells a specified security at a specified time instant. Potters *et al.* (2003)
Ask size	The number of shares in lots that are quoted at the asking price. Potters *et al.* (2003)
Turnover ratio	Measures a firm's trading frequency. Amihud *et al.* (1986)
Price change	The natural logarithm of the ratio of the mid quote price subsequent to the trade, to the mid quoted price before the trade. Lillo *et al.* (2003)
Spread	The difference between the ask and the bid. Glosten *et al.* (1988)
Mid quote price	The mid quote price is the midpoint of the ask and the bid. Potters *et al.* (2003)
Average value traded	The moving mean of the individual share values exchanged over a specified duration. Harvey *et al.* (2016)
Volatility	The standard deviation of natural logged returns. Micciche *et al.* (2002)
Momentum	The ratio of the prevailing price to the lag one prevailing price. Jegadeesh *et al.* (1993)
KL divergence	The degree of change in the empirical distribution of counts or sizes of orders, which are ordered according to price. Kullback *et al.* (1951)
Order sign	Classification of order as either buyer initiated or seller initiated. Lee *et al.* (1991)
Signed volume	The signed order is the product of the order size and the order sign. Bouchaud *et al.* (2017)
Order imbalance	Order imbalance is the sum of signed orders over a specified duration. Lillo *et al.* (2003)
Liquidity	The slope of the linearized version of the price impact model. Lillo *et al.* (2003)
Execution time	Execution time is the time latency between order match instant and the appearance of the subsequent event. Moro *et al.* (2009)
PIN	The probability of informed trading. Easely *et al.* (1996)
Normalized volume	The normalized volume is measured as the size of each order scaled by the mean volume traded. Lillo *et al.* (2003)

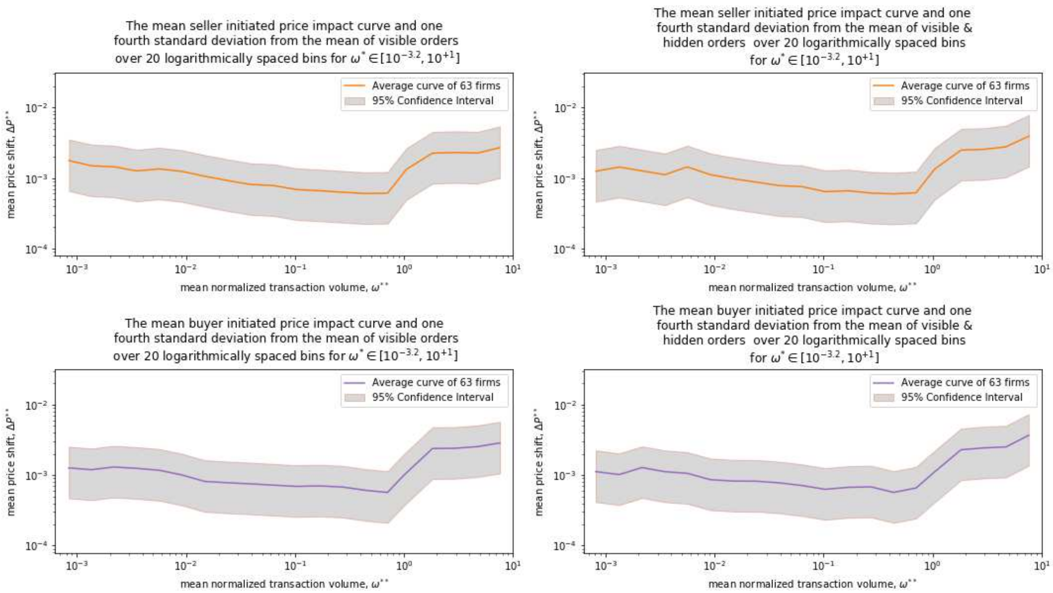

Fig. A.21. The mean price impact curves and the 95% confidence interval of a constant standard deviation from the mean of single trades of 63 liquid firms. The curves to the left are derived from the data of visible orders; meanwhile, the curves to the right are derived from the data of visible & hidden orders.

Johannesburg Stock Exchange (JSE). *Approach* : Lillo *et al.* (2003)& Harvey *et al.* (2016)

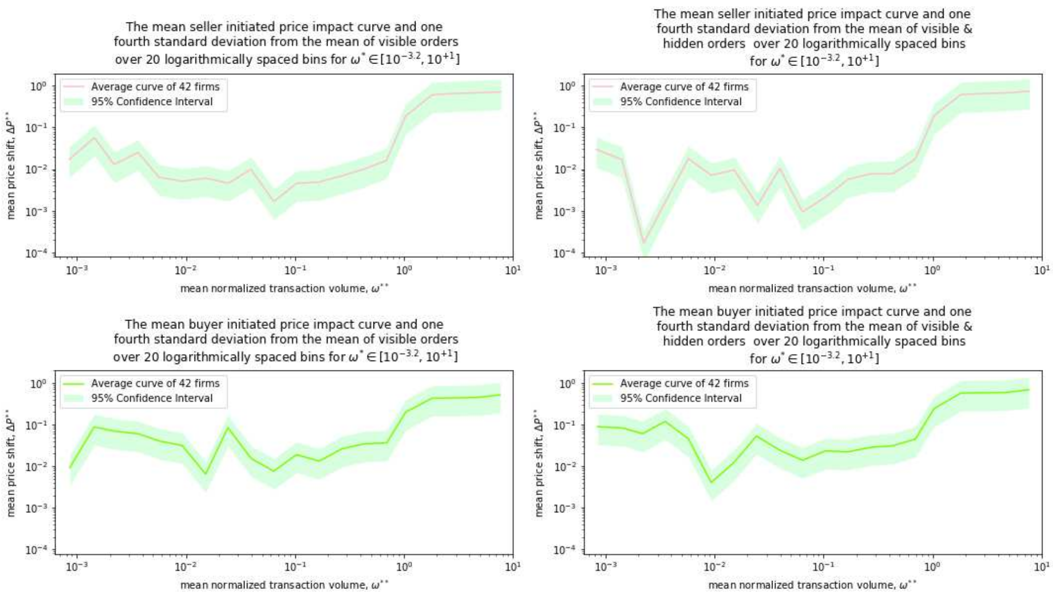

Fig. A.22. The mean price impact curves and the 95% confidence interval of a constant standard deviation from the mean of single trades of 42 liquid firms. The curves to the left are derived from the data of visible orders; meanwhile, the curves to the right are derived from the data of visible & hidden orders.

Bolsa de Valores de Sao Paulo (BOVESPA). *Approach* : Lillo *et al.* (2003)& Harvey *et al.* (2016)

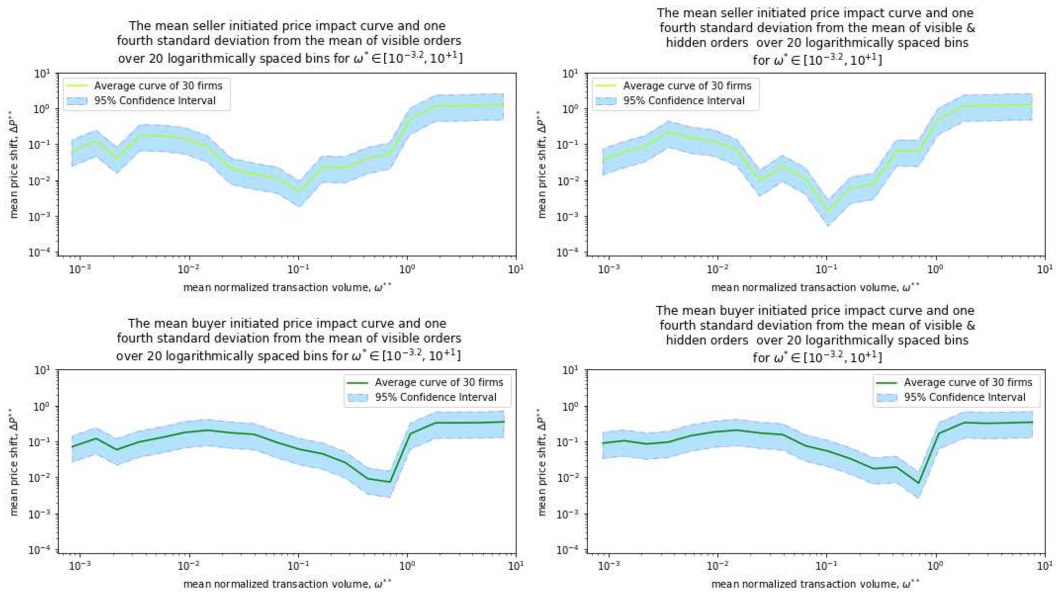

Fig. A.23. The mean price impact curves and the 95% confidence interval of a constant standard deviation from the mean of single trades of 30 liquid firms. The curves to the left are derived from the data of visible orders; meanwhile, the curves to the right are derived from the data of visible & hidden orders.

Moscow Exchange (MOEX). *Approach* : Lillo *et al.* (2003)& Harvey *et al.* (2016)

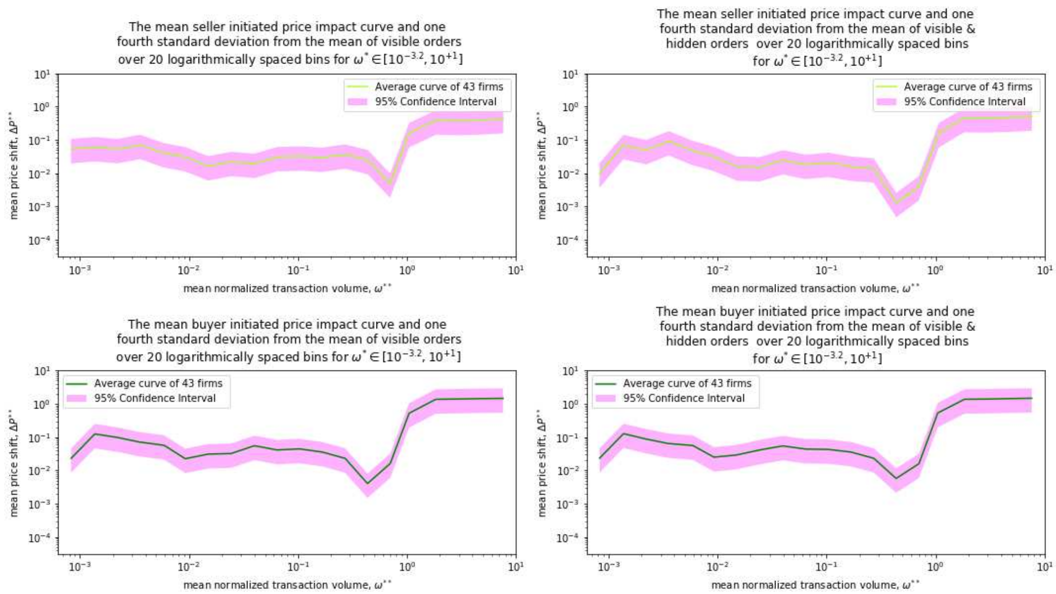

Fig. A.24. The mean price impact curves and the 95% confidence interval of a constant standard deviation from the mean of single trades of 43 liquid firms. The curves to the left are derived from the data of visible orders; meanwhile, the curves to the right are derived from the data of visible & hidden orders.

National Stock Exchange of India (NSE). *Approach* : Lillo *et al.* (2003)& Harvey *et al.* (2016)

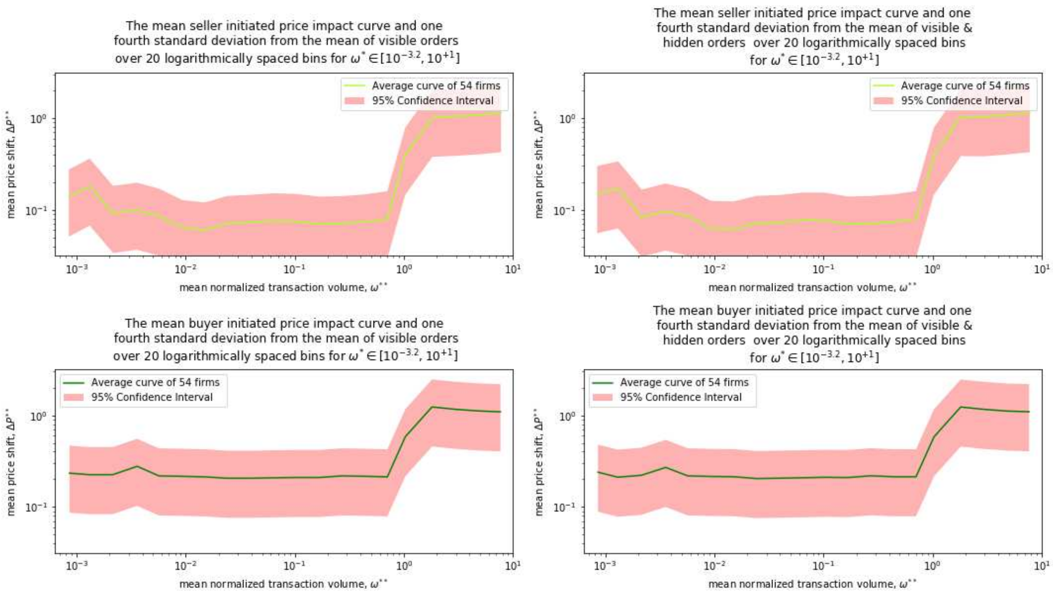

Fig. A.25. The mean price impact curves and the 95% confidence interval of a constant standard deviation from the mean of single trades of 54 liquid firms. The curves to the left are derived from the data of visible orders; meanwhile, the curves to the right are derived from the data of visible & hidden orders.

Shanghai Stock Exchange (SSE). *Approach* : Lillo *et al.* (2003)& Harvey *et al.* (2016)

Acknowledgements

We would like to bring to the attention of the reader that the research in case would not have been possible without the invaluable contributions of the following participants: Wycliffe Ouma, Warren Carlson and Marike Kluyts. We thank Wycliffe Ouma for providing access to the data portal. We thank Warren Carlson for suggestions on the technical approach. We are grateful to Marike Kluyts for the incredible editorial and review of the research.

Disclaimers

The views expressed in this paper are those of the authors; therefore, fellow researchers are not necessarily advised to follow these views. Writing errors, if any exist, are the authors' sole responsibility.