Journal of Forecasting, *J. Forecast.* (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/for.2422

Identifying Expensive Trades by Monitoring the Limit Order Book

BENOIT DETOLLENAERE AND CATHERINE D'HONDT*

ABSTRACT
This paper provides an analysis of the impact of microstructure variables on the transaction costs for split orders on 171 Euronext large cap stocks. First, using the adaptive Lasso selection method, we conduct an exploratory study to identify which microstructure variables best explain the transaction costs of split orders. Then, we propose an ordinal logistic model to classify ex ante transaction costs into buckets. Our empirical work demonstrates that our model performs very well both in-sample and out-of-sample. All the findings show that microstructure information is of great importance for any investor who attempts to manage the execution of split orders throughout the day. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS split orders; transaction costs; limit order book; Lasso

INTRODUCTION

Trading is the execution of investment decisions. It is well established that the quality of execution is as relevant as the decision itself because of transaction costs that burden the total portfolio performance, especially for active management. These transaction costs are generally divided into explicit and implicit costs. Explicit costs are direct costs of trading (fees, commissions and taxes), whereas implicit costs mainly depend on the trade characteristics relative to the prevailing market conditions.

Investors are often more comfortable with explicit costs that can be determined before the execution of the trade. Implicit costs represent an invisible part that cannot be measured ex ante because they are included in the trade price. These costs are usually decomposed into spread,[1] market impact and opportunity costs.[2] Among them, the market impact, which can be defined as the cost incurred for consuming the liquidity available in the order book beyond the best opposite quote (BOQ), is undoubtedly the cost that raises the most concerns. As it results in a price shift due to the trade size, it can be very expensive for large trades, in particular on illiquid securities.

To manage market impact, traders often tend to split their large orders into smaller pieces that are executed over time. This splitting strategy may be implemented by the trader himself or with the use of a dedicated algorithm. Unfortunately, the contribution of market impact to total implicit costs is not independent of opportunity costs, especially the market timing cost that arises because of potential adverse price movement during the time required to complete the trade. This tricky issue is referred to as the so-called trader's dilemma. This dilemma has been mathematically formulated and widely studied by many authors (Bertsimas and Lo 1998, Almgren and Chriss, 1999, 2001; Kissell *et al.*, 2004). Similarly to the mean–variance optimization introduced by Markowitz (1952), they apply optimization techniques to solve the trader's dilemma. Bertsimas and Lo (1998) derive execution strategies which optimally adapt to price variations and market conditions. Almgren and Chriss (1999) propose a mean–variance optimization framework that aims at determining an optimal trade schedule that balances market impact and market timing risk based on traders' appetite for risk. Kissell *et al.* (2003) build their techniques on Almgren and Chriss (1999) but consider risk rather than variance in their optimization process.[3] This stream of literature reveals a constant concern to strike a balance between the cost of immediacy and the risk of trading at a worse price and offers a strong argument for examining sequences of incrementally executed orders rather than on orders on a stand-alone basis. This interest for sequences of orders has been recently confirmed in O'Hara (2015) who provides a big picture of how markets have evolved and trading has changed. With algorithms that chop a parent order into child orders, the resulting split orders are not independent and the sequences of trades become informative. Such sequences of split orders are called 'hidden orders' by some authors (e.g. Moro *et al.*, 2009) because traders attempt to hide the true size of their orders. In this paper, we will simply refer to split orders.

*Correspondence to: Catherine D'Hondt, Louvain School of Management – Université catholique de Louvain, Chaussée de Binche 151, 7000 Mons, Belgium. E-mail: catherine.dhondt@uclouvain.be

[1] The bid–ask spread is a compensation for the supply of liquidity.

[2] The opportunity costs are due to the adverse price movement that takes place between the trade decision and the trade itself. They are made of three different components: operational opportunity costs, market timing opportunity costs and missed trade opportunity costs. The first costs arise when the delay required to trade is operational; the second costs are due to the market timing under the control of the broker; the missed trade opportunity costs occur when the trader is not able to fully fill his order.

[3] In fact, minimizing with the risk term is more intuitive but introduces complexities in the optimization.

Although their practical importance is well known and their use is widespread with algorithmic trading, only a few studies have investigated split orders, and in particular the determinants of their transaction costs. The main reason is that it requires the identity of the trading accounts behind orders and such information is often difficult to obtain because of confidentiality. The few empirical papers on the topic focus on institutional orders made available to the authors by large financial institutions. Among them, the reference papers are Chan and Lakonishok (1995) and Keim and Madhavan (1997), which both analyze the relationship between transaction costs and trade difficulty variables, as well as stock-specific information and investment styles, respectively. Such studies are of great interest but they only refer to a fraction of total market activity and result in a small sample size (e.g. only large orders from some institutions). Besides, as they use brokerage data, they lack most of the market data and especially limit order book data. More recently, some authors (Lillo *et al.*, 2008; Vaglica *et al.*, 2008; Moro *et al.*, 2009) characterized split orders identified among transactions from market data but, to the best of our knowledge, none investigates their transaction costs. Nevertheless, addressing this issue would contribute to a better understanding of split orders and, to a larger extent, of both traders' behavior and market functioning. It could also lead to concrete applications in the financial industry. This is particularly true for investors who are concerned with transaction costs forecasting.

In this paper, we fill this gap in two ways. First, we conduct an exploratory analysis to identify the determinants of the implicit transaction costs of split orders on 171 Euronext large cap stocks. Specifically, using the adaptive Lasso selection method of Zou (2006) in a linear framework, we select the set of variables that best explain transaction costs among a large panel including microstructure variables. Second, as it is hard to forecast transaction costs with accuracy, we propose an alternative ordinal logistic model that aims at classifying ex ante split orders into 'buckets' that represent different levels of transaction costs. We then use this model to check whether investors could effectively predict the magnitude of their transaction costs based on microstructure information.

The main findings are the following. First, regarding our exploratory analysis, the results suggest that the order flow imbalance and the proportion of canceled limit orders 15 minutes before execution are significant determinants of the transaction costs of split orders. The shape of the limit order book seems also to determine transaction costs but to a lesser extent. The size of the sequence of split orders as well as its degree of aggressiveness are positively related to transaction costs. As for the detection of hidden liquidity, this plays a role in both buy and sell split orders up to 15 minutes. Second, our logit model built on microstructure variables performs very well in forecasting the level of transaction costs of split orders up to 30 minutes. The results are conclusive both in-sample and out-of-sample. We are able to predict, on average, more than 50% of the most expensive trades. We reach an average percentage of correctly classified orders of about 40% as well as a relatively low percentage of serious misclassification of about 25%. All these findings provide evidence that microstructure variables are of great importance in managing the implicit transaction costs of split orders.

The remainder of this paper is organized as follows. The next section describes our dataset and the sample characteristics. The third section presents the transaction cost measures and the set of explanatory variables under scrutiny. The fourth section describes the methodology. The fifth section reports the results of our empirical work. The sixth section concludes.

DATA AND SAMPLE

We use tick data from Euronext.[4] This market operates an electronic order-driven system. Whenever an order is submitted at a price that is equal to or better than the BOQ, trades take place. Orders that are not immediately fully executed are recorded in the order book, for which the system enforces price–time priority. All retail and institutional order flow is routed through market members, who can act either as brokers when submitting orders on behalf of clients (institutional and individual) or as dealers when trading on their own account.[5] Trading for most of the liquid stocks occurs in various phases throughout the day. Continuous trading starts just after the opening call auction around 9:00 am and ends at 5:25 pm. The market is quite transparent, as it diffuses at any time all the quotes on each side of the order book, with the associated displayed depth. Transactions (price and quantity) are also disclosed. However, traders can use hidden orders, for which only some fraction of the total quantity is disclosed to other market participants.[6] In addition, the market is anonymous: traders do not know who places orders in the book (pre-trade anonymity), nor do they know who is involved in trades (post-trade anonymity).

Our sample contains 171 large cap stocks[7] and 61 trading days from 1 February to 30 April 2006. This sample period has the advantage of being representative of the total market activity before the fragmentation of the trading

[4] Euronext was the European regulated cash market of the NYSE Euronext group. The latter is the holding company created in 2007 by the combination of NYSE Group, Inc. and Euronext NV. Some years earlier, Euronext NV was created to bring together major marketplaces across Europe and become the first pan-European market. Euronext was, at that time, referred to as the European regulated cash market of the NYSE Euronext group, with nearly 1750 firms listed across four countries (Belgium, France, Portugal and The Netherlands). On 13 November 2013 Intercontinental Exchange (ICE) acquired NYSE Euronext. In 2014 Euronext became again a stand-alone company through an initial public offering.

[5] For some illiquid stocks, they can also act as liquidity providers. A liquidity provider has an agreement with the exchange authorities or with listed companies themselves in order to supply liquidity during the whole trading session.

[6] Accordingly, hidden quantity becomes visible when the displayed quantity is executed and it keeps price priority but loses time priority.

[7] Our dataset contains 701 stocks that are traded during the continuous session on Euronext from February 2006 to April 2006. These 171 stocks correspond to the stocks with a market capitalization above the third quartile of the market capitalization distribution of the entire sample.

landscape and the emergence of MTF due to MiFID.[8] The dataset is the same as used in De Winne and D'Hondt (2007). For each stock, it includes common public market data about both orders and trades as well as additional private information. Based on them, we are able to get accurate order book data, including aggregate both displayed and hidden quantities available at each quote. In particular, we also have the ID code[9] for the market member placing the order. We also know for which account type he is trading, which allows us to distinguish between client trading and proprietary trading.

To identify split orders, i.e. large orders broken into pieces and executed incrementally, we follow Moro *et al.* (2009). Nevertheless, our approach differs in three ways. First, Moro *et al.* (2009) have access to ID codes but ignore when the market member acts as a broker or on his own account. As a consequence, they have to infer sequences of orders based on a hidden Markov algorithm. In our dataset, the proprietary/client dummy permits us to directly identify with accuracy which orders belong to a given sequence of orders submitted by the same market member for proprietary trading. This implies that we only keep in our final sample proprietary orders. Second, whereas Moro *et al.* (2009) analyze both non-marketable orders and marketable orders,[10] we focus on marketable orders only because we do not consider orders that provide liquidity. Implicit transaction costs are essentially an issue for liquidity takers. Finally, we do not consider order sequences with an execution time exceeding 30 minutes since the impact of microstructure variables is expected to be short lived according to the literature. For example, Cao *et al.* (2009) report an effect of the limit order book imbalance on returns that does not exceed 5 minutes. In a nutshell, our approach implies that our final sample is therefore made of split orders defined as sequences of proprietary orders whose completion time does not exceed 30 minutes.

TRANSACTION COST MEASURE AND DETERMINANTS

Transaction costs

To measure implicit transaction costs, we use the benchmark comparison method, which is based on the signed difference between the average execution price of the trade and a benchmark price. Given our rich dataset, we take as benchmark price the BOQ prevailing at the submission time of the first order of a sequence (BOQ_1). Accordingly, the transaction costs of a sequence j of n split orders i expressed in basis points ($\mathrm{TC}_j^{\text{buy/sell}}$) is computed as follows:

$$\mathrm{TC}_j^{\text{buy/sell}} = \left| \left[\frac{\frac{\sum Q_i * \mathrm{AEP}_i}{\sum_i^n Qi}}{\mathrm{BOQ}_1} - 1 \right] \right| * 10{,}000 \text{ for } i = (1, \ldots, n) \tag{1}$$

where AEP_i and Q_i are the average execution price and the size of each order i, respectively.[11]

This measure allows us to investigate transaction cost determinants related to both liquidity and short-term price evolution. Indeed, it depends on liquidity as the average execution price (AEP_i) is inversely proportional to the liquidity available. But it is also related to the price evolution of the stock during the execution of the split orders, as this average execution price also captures the market price evolution through time.

We report some descriptive statistics on the sequences of split orders in Table I. Consistent with the literature, transaction costs increase with the traded volume. We also observe that larger sequences are executed over a longer period of time, which confirms that traders split larger orders into smaller pieces to reduce their market impact cost. The transaction costs of buy orders are also slightly larger than those of sell orders, while the volume, size and the duration display similar profiles. One plausible explanation is that the market rose by almost 4% during the sample period, which made buy split orders more costly than sell split orders, on average.

Transaction cost determinants

A major contribution of this paper is that we investigate the effect of a large set of predictors on transaction costs of split orders. In this section, we review the different variables under scrutiny. As just mentioned, we investigate variables related to liquidity and to the short-term price evolution. Specifically, we classify our variables into three groups:[12] price movement variables, liquidity determinants and other determinants. The last group encompasses variables that do not clearly belong to the first two categories.

[8] MTF stands for Multilateral Trading Facilities and MiFID for Markets in Financial Instruments Directive.

[9] Actually, these ID codes are numerical in order to ensure market members' anonymity but allow us to isolate the whole set of orders or trades associated with a given member from the other orders and trades in the sample.

[10] A marketable order is a market order or a limit order that executes immediately. A non-marketable order is a limit order that does not execute immediately and is then registered in the limit order book for later potential execution.

[11] Since we focus on sequences of split orders with a time to execution that does not exceed 30 minutes, there is no need to adjust our measure of transaction costs for any market movement as suggested in Chan *et al.* (1995).

[12] Note that some variables may belong to more than one category. Despite the limits, this classification has the merit of structuring things.

Table I. Descriptive statistics for the sequences of split orders.

	5 minutes	15 minutes	30 minutes
Panel A: Buy orders	N (79,272)	N (48,652)	N (29,712)
Volume	163,180.43	212,090.92	256,748.85
Size	3.11	4.59	6.11
Time	00:01:45	00:09:12	00:21:30
Cost	0.86	1.41	1.26
Panel B: Sell orders	N (78,369)	N (48,735)	N (30,357)
Volume	155,643.11	202,517.96	253,754.55
Size	3.11	4.58	6.04
Time	00:01:44	00:09:12	00:21:27
Cost	0.72	1.23	1.24

Note: Cross-sectional statistics on the sequences of split orders are reported for the 171 stocks, regarding both their direction (buy/sell) and their approximate time to execution. The sequences that belong to the 5-minute group have a duration below 5 minutes. The 15-minute group is made of the sequences with a duration between 5 and 15 minutes, while the 30-minute group contains the sequences with a duration between 15 and 30 minutes. N refers to the number of sequences. 'Volume' refers to the average volume expressed in units of currency. 'Size' is the average number of orders executed within a sequence. 'Time' is the average duration for the sequences. 'Cost' is the average transaction cost for the sequences expressed in basis points.

Price movement variables

Order flow and shape of the limit order book. In order-driven markets, the order book provides a lot of information and some of them are strong candidates for explaining the price evolution component of transaction costs. Some authors (e.g. Chordia *et al.*, 2002; Chordia and Subrahmanyam, 2004; Lee *et al.*, 2004) show that order flow imbalance can be used to predict returns. Cao *et al.* (2009) also outline that the imbalance between the bid and the ask side of the order book could help predict returns. More recently, Tombeur and Wuyts (2011) investigate whether the limit order book could help to predict returns. They analyze two categories of information: information on the order flow and the shape of the limit order book. In this paper, we replicate and adapt some of their measures.

For the order flow imbalance, we first classify orders into five categories according to their aggressiveness. Category 1 refers to a marketable order that consumes more than the displayed liquidity at the BOQ. Category 2 refers to a marketable order that consumes less than the displayed liquidity at the BOQ. Category 3 is made of limit orders that improve the spread. Category 4 refers to limit orders at the best quote on the same market side. Category 5 encompasses limit orders beyond the best quote on the same market side. Then, we compute the imbalance between the buy volume (B_{jt}) and the sell volume (S_{jt}) from each order category j during the 15-minute interval preceding each sequence of split orders. Order imbalance for a sequence of split orders starting at time T (Oimb_{jT}) is then computed for the five different categories of orders submitted during the 15-minute interval $[T-15, T]$ as

$$\mathrm{Oimb}_{jT} = \sum_{t=T-15}^{T} \frac{B_{jt} - S_{jt}}{B_{jt} + S_{jt}} \quad \text{for } j = (1, \ldots, 5) \tag{2}$$

Tombeur and Wuyts (2011) find that order flow imbalances for both market orders and limit orders are informative for short-term returns. They suggest that market orders are highly autocorrelated and generate short-term price pressure, while limit orders contain some information on the fundamental value of the security. Consistent with their findings, we expect a positive (negative) relationship between our measures of order imbalance and the transaction costs of buy (sell) split orders since the short-term return is expected to increase (decrease) with them.

We also compute the trade imbalance (Timb_T) during the same 15-minute interval $[T-15, T]$. It differs from the previous measures since it is based on trades and thus only considers executed volumes, and hence volumes from category 1 and category 2 orders for buy and sell, respectively:

$$\mathrm{Timb}_T = \sum_{t=T-15}^{T} \frac{\mathrm{BuyVol}_t - \mathrm{SellVol}_t}{\mathrm{BuyVol}_t + \mathrm{SellVol}_t} \tag{3}$$

where BuyVol_t and $\mathrm{SellVol}_t$ stand for trade volume at time t from executed orders of category 1 and 2.

Building on Tombeur and Wuyts (2011), we also summarize the shape of the limit order book at time t, i.e. just before the inception of each sequence of split orders, into two measures: depth imbalance ($\mathrm{Dimb}_{i,t}$) and height imbalance ($\mathrm{Himb}_{i,t}$).

The depth imbalance ($\text{Dimb}_{i,t}$) at time t is computed at each level i of the limit order book as follows:

$$\text{Dimb}_{i,t} = \frac{D_{i,t}^{\text{bid}} - D_{i,t}^{\text{ask}}}{D_{i,t}^{\text{bid}} + D_{i,t}^{\text{ask}}} \text{ for } i \in [\![1 \ldots 5]\!] \tag{4}$$

where $D_{i,t}^{\text{bid}}$ and $D_{i,t}^{\text{ask}}$ are the number of displayed shares available at time t at the ith limit of the order book on the bid side and on the ask side, respectively. Depth imbalance is expected to be positively (negatively) related to transaction costs for buy (sell) split orders. The reason is twofold. First, Tombeur and Wuyts (2011) report that an increase (decrease) in depth imbalance indicates an increase (decrease) in returns, since a deeper order book on the bid (ask) reveals a buying (selling) interest. Second, an increase in depth imbalance also means that there is an increase (decrease) of the liquidity available on the bid (ask) side relative to the ask (bid) side.

Finally, the imbalance in height ($\text{Himb}_{i,t}$) is computed exactly as in Tombeur and Wuyts (2011):

$$\text{Himb}_{i,t} = \frac{H_{i,t}^{\text{ask}} - H_{i,t}^{\text{bid}}}{H_{i,t}^{\text{ask}} + H_{i,t}^{\text{bid}}} \text{ for } i \in [\![2 \ldots 5]\!] \tag{5}$$

where $\text{Himb}_{i,t}$ is the imbalance between the absolute value of the difference between the limit i and limit $i-1$ at time t on the same side of the order book.[13] This measure is expected to be positively (negatively) related to buy (sell) split orders transaction costs for two reasons. First, as reported by Tombeur and Wuyts (2011), if prices are more crowded on the bid side than on the ask side, it may reveal a buying interest and therefore generate a positive price movement. The same holds for the ask side. Second, a decrease (increase) in the height imbalance means that the order book is more crowded on the ask (bid) side and, as a result, walking down the order book becomes more costly for buy (sell) orders. In other words, the height imbalance also measures the 'cost of jumping' from one step to a lower step in the order book.

Fleeting orders. Hasbrouck and Saar (2009) define these as non-marketable limit orders that are rapidly canceled after submission. These authors show that there is an increase in the cancellation of limit orders when market prices move away from the limit order price. They argue that informed investors attempt to resubmit more aggressive orders to chase the price that moves away. They also find weaker evidence that the cancellation of limit orders is more likely to occur when the BOQ becomes closer to the limit order price. They explain this phenomenon by the fact that investors switch to market orders because the cost of immediacy decreases.

In line with their results, we expect that an increase in limit orders cancellations just before a sequence of split orders will impact transaction costs. An increase in the aggressiveness on the bid (ask) side subsequent to an increase of the cancellation on the same side (bid/ask side) should result in an increase (decrease) in transaction costs for buy orders and the opposite for sell orders.

In order to measure order cancellations, we compute the ratio of the number of canceled buy/sell orders to the total number of orders submitted during the 15 minutes preceding each sequence of split orders. The ratio of buy/sell canceled orders for a sequence of split orders starting at time T ($\text{Cancel}_T^{\text{buy/sell}}$) is then computed during the 15-minute interval $[T-15, T]$ as

$$\text{Cancel}_T^{\text{buy}} = \sum_{t=T-15}^{T} \frac{\text{NC}_t^{\text{buy}}}{N_t^{\text{buy}}} \tag{6}$$

$$\text{Cancel}_T^{\text{sell}} = \sum_{t=T-15}^{T} \frac{\text{NC}_t^{\text{sell}}}{N_t^{\text{sell}}} \tag{7}$$

where $\text{NC}_t^{\text{buy/sell}}$ and $N_t^{\text{buy/sell}}$ stands for the number of canceled orders and the total number of submitted orders at time t for buy and sell orders, respectively.

Liquidity variables

Consistent with Keim and Madhavan (1997), we use the log of market capitalization as a proxy of liquidity. The higher the market capitalization, the lower are the transaction costs. Following Bikker *et al.* (2007), we also assess liquidity with the average daily volume computed over the entire sample period.

Some authors (Chan and Lakonishok, 1997; Keim and Madhavan, 1997) outline that the market place matters for transaction costs. More specifically, they examine the difference in transaction costs between exchange-listed

[13] For example, $H_{2,t}^{\text{ask}}$ is the absolute value of the difference between the best ask and the second best ask. Note that we do not define the first step since the only way to define the height at this step is to use the midpoint and the resulting height imbalance would be 0.

stocks and comparable Nasdaq stocks. Our sample of split orders is executed on four different market places (Paris, Amsterdam, Brussels and Lisbon), which have similar liquidity profiles. The inclusion of variables controlling for the market place in our list of determinants of transaction costs is thus subject to debate. On the one hand, Bikker *et al.* (2007) consider Euronext Amsterdam, Brussels, Paris and Lisbon as one single exchange, because of the similarities in trading systems. On the other hand, Domowitz *et al.* (2001) report significant differences in transaction costs across the four Euronext venues.[14] To be on the safe side, we include dummies to control for the trading venue.

Many studies provide empirical evidence of intraday seasonality of liquidity on different trading platforms.[15] The implication is that transaction costs vary throughout the day and that any model attempting to predict transaction costs must take this intraday seasonality pattern of liquidity into account. Building on Bikker *et al.* (2008), we slice the trading day into five periods and create dummies to control for intraday seasonality. On Euronext, we expect an inverse J-curve for transaction costs as illustrated in Gomber *et al.* (2004). Finally, we also include dummies for a potential 'day-of-the-week' effect as in Bikker *et al.* (2008).

Other variables

To control for trade difficulty, we first follow Bikker *et al.* (2008) and take the square root of the trade size relative to the average daily volume of the stock. Then, as in Moro *et al.* (2009), we include variables that characterize the execution of split orders. These are composed of the time to execute the sequence of split orders (T), the number of split orders (N) and a variable (Vol_1) reflecting the degree of aggressiveness of the execution. The latter is the ratio of aggressive orders consuming the depth beyond the BOQ relative to the entire volume of the sequence.[16]

To capture the impact of undisclosed liquidity on transaction costs, we include a dummy for the detection of hidden liquidity at the BOQ based on the replenishment event described in Detollenaere (2015). This dummy variable takes a value of one if undisclosed depth at the BOQ has been detected just before the sequence of split orders is initiated and zero otherwise.

We also include the relative spread prevailing at the beginning of the sequence to control for the mean reversion in the bid–ask spread documented by Biais *et al.* (1995). They report that orders within the quotes tend to be placed when the spread is large, while trades occur more frequently when the spread is tight. These effects generate mean reversion in the bid–ask spread. In simple terms, a large spread is more likely to decrease and return to a more normal level. We therefore expect a decrease (increase) in transaction costs when the spread is large (tight) since the BOQ is more likely to get closer (shift away) when the spread is large (small).

Finally, volatility is likely to impact transaction costs and must therefore be analyzed. Our proxy of volatility is the high–low of the trade prices during the 15-minute interval preceding each sequence of split orders, scaled by the last trade price of the interval. Among the different possible volatility measures, this one enables us to avoid any computational issue resulting from no trading activity periods that could arise during the day. Table II provides a complete list of the variables, with their abbreviations and definitions.

METHODOLOGY

The adaptive Lasso selection model

Our first objective is an exploratory analysis to identify the determinants of the implicit transaction costs of split orders. For this purpose, we use the adaptive Lasso selection method of Zou (2006) to select the set of variables that best explain transaction costs among the large panel of variables described above.

Referring to Tibshirani (1996), our motivation for using a model selection such as the Lasso is threefold. Firstly, it allows from a large number of variables to determine a smaller subset that exhibits strong effects, which in turn facilitates interpretation of the explanatory variables parameter estimates. Secondly, it improves prediction accuracy by setting to zero some coefficients, thereby sacrificing a little bias to reduce the variance of the predicted value. Lastly, the Lasso method facilitates the deployment of any model in practice as it quickly generates the best model with fewer variables instead of testing 2^p models for p variables. This point is particularly relevant in our situation, and more generally for any predicting model at a high frequency level, because it permits a rapid re-estimation of the best model and a lower amount of computation to feed the model.[17]

[14] Note that their study covers a sample period anterior to the merger of the four venues. But Nielsson (2009), based on different liquidity measures, report asymmetric liquidity gains from the merger, where the positive effects are concentrated among large firms and firms with foreign sales. This suggests that some liquidity disparities remain across the different venues of Euronext.

[15] Some examples are Abhyankar *et al.* (2003) for the London Stock Exchange, Foster and Viswanathan (1990) for the New York Stock Exchange, Chan *et al.* (1995) for NASDAQ stocks and Hamao and Hasbrouck (1995) for the Tokyo Stock Exchange.

[16] This measure differs slightly from the fraction of market orders variable presented in Moro *et al.* (2009), which is the traded volume of split market orders divided by the total volume of split orders.

[17] Some subset selection procedures present the same practical advantages (e.g. backward selection procedure) but, as outlined by Tibshirani (1996), they are extremely variable because they operate in a discrete process in which regressors are either retained or dropped from the model. Consequently, small changes in the data can result in very different models being selected. Besides, the Lasso is supported by much theoretical work (Donoho *et al.*, 1995; Donoho and Elad, 2003; Donoho, 2006; Meinshausen and Bühlmann, 2006; Zou, 2006).

Table II. Potential determinants of transaction costs and their definitions

Determinants	Definition
Price movement variables	
Oimb_j	Scaled order imbalance for order category j
Timb	Scaled trade imbalance for order category 1 and 2
Dimb_i	Scaled depth imbalance at step i of the limit order book
Himb_i	Imbalance in height between step i and step $i-1$ in the book
$\text{Cancel}^{\text{buy}}$	Number of canceled buy orders relative to the total number of buy orders
$\text{Cancel}^{\text{sell}}$	Number of canceled sell orders relative to the total number of sell orders
Liquidity variables	
Logcapi	Logarithm of market capitalization
ADV	Average daily volume over the sample period
FR	0/1 variable for split orders executed on the Paris Stock Exchange
PT	0/1 variable for split orders executed on the Lisbon Stock Exchange
BE	0/1 variable for split orders executed on the Brussels Stock Exchange
(NL)	0/1 variable for split orders executed on the Amsterdam Stock Exchange
Opening	0/1 variable for split orders submitted between 9:00 am and 10:30 am
Morning	0/1 variable for split orders submitted between 10:30 am and 12:30 am
(Midday)	0/1 variable for split orders submitted between 12:30 am and 3:00 pm
Afternoon	0/1 variable for split orders submitted between 3:00 pm and 4:00 pm
Closing	0/1 variable for split orders submitted between 4:00 pm and 5:30 pm
Mon	0/1 variable for split orders executed on Monday
Tue	0/1 variable for split orders executed on Tuesday
(Wed)	0/1 variable for split orders executed on Wednesday
Thu	0/1 variable for split orders executed on Thursday
Fri	0/1 variable for split orders executed on Friday
Other variables	
Order	Square root of split orders size relative to the average daily volume
Time	Time to execute the sequence of split orders
N	Number of orders in a sequence of split orders
Vol_1	Volume of aggressive orders beyond the BOQ relative to the total volume of split orders
Hidden	0/1 variable for detected hidden liquidity at the BOQ at the inception of a sequence of split orders
Rspread	Relative spread at the inception of a sequence of split orders
HL	High–low of trade prices during the 15-minute interval before the inception of a sequence of split orders

Note: The dummy variables in parentheses have not been included in the estimated models to avoid exact collinearity. Split orders initiated before 9:15 am are excluded from the sample since at least 15 minutes of continuous session are required to compute some of our explanatory variables.

The Lasso (least absolute shrinkage) method of Tibshirani (1996) is a constrained ordinary least squares (OLS) regression, where the sum of the absolute values of the regressors cannot be higher than a specified parameter. Consider X=$(x_1, x_2, \ldots, x_n)$ the matrix of regressors and y the dependent variable, where the x_i are standardized. Then the $B = (\beta_1, \beta_2, \ldots, \beta_n)$ Lasso regression coefficients, for a given parameter t, are the solution to the following constrained optimization problem:

$$\text{Minimize } \parallel y - X\beta \parallel^2 \text{ subject to } \sum_{j=1}^{m} \mid \beta_j \mid \leq t \tag{8}$$

Basically, we can view the Lasso as selecting a subset of the regression coefficients for each Lasso parameter. Through an increase in discrete steps of the Lasso parameter, we obtain a sequence of coefficients where the non-zero coefficients are the selected parameters.

Zou (2006) proposes the adaptive Lasso as an improvement on Tibshirani (1996). Its purpose is to reduce the selection bias by allocating higher penalty to zero coefficients and lower penalty to non-zero coefficients. Assuming that we can find a suitable estimator $\hat{\beta}$ of the parameters in the true model and a weight vector defined by $w = \frac{1}{|\beta|^{\zeta}}$ (where $\zeta \geq 0$), the adaptive Lasso is defined as follows:

$$\text{Minimize } \parallel y - X\beta \parallel^2 \text{ subject to } \sum_{j=1}^{m} \mid w_i \beta_j \mid \leq t \tag{9}$$

where w_i are the adaptive weights for the different parameters.

We implement the above model as follows. First, as suggested by Zou (2006), we use the solution to the unconstrained least squares model as the estimator for $\hat{\beta}$.[18] To solve the constrained least squares problem, we use Efron *et al.*'s (2004) algorithm. The latter is roughly a stepwise procedure with a single addition to or deletion from the set of non-zero regression coefficients at any step. Then, we take the Bayesian information criterion (BIC) as the criterion to choose among the models at each step of the Lasso algorithm. This criterion is partly based on the likelihood function as the Akaike information criterion (AIC) but penalizes more the complexity of the model as it gives more penalty to the number of parameters in the model. Concretely, we apply this Lasso selection procedure to the three panels of sequences of split orders presented in Table I with different durations (5, 15 and 30 minutes) in each direction (buy/sell).

The forecasting model

Our second objective aims at classifying ex ante split orders into 'buckets' that represent different levels of transaction costs. For this purpose, we propose the following alternative ordered logit model.

Let Y_i be the ordered categorical dependent variable for observation i that takes one of the integer values from 1 to J, where J is the total number of categories. The random component starts with an unobserved continuous variable Y_i^*, which follows the standard logistic distribution with a parameter μ_i:

$$Y_i^* \sim \text{ Logit } (y_i^* \mid \mu_i) \tag{10}$$

to which we assign different thresholds:

$$Y_{i=j} \ if \ \tau_{j-1} \le Y_i^* \le \tau_j (for j = 1, \dots, J) \tag{11}$$

where τ_j (for $j = 0, \dots, J$) are the threshold parameters with $\tau_j \le \tau_m$ for all $j \le m$ and $\tau_0 = -\infty$ and $\tau_j = \infty$.

Given the parameters τ_j and β, and the regressors x_i, the non-random component is as follows:

$$\Pr(Y \le j) = \Pr(Y^* \le \tau_j) = \frac{\exp(\tau_j - x_i \beta)}{1 + \exp(\tau_j - x_i \beta)} \tag{12}$$

and the predicted probabilities π_j for each category j follow:

$$\pi_j = \frac{\exp(\tau_j - x_i \beta)}{1 + \exp(\tau_j - x_i \beta)} - \frac{\exp(\tau_{j-1} - x_i \beta)}{1 + \exp(\tau_{j-1} - x_i \beta)} \tag{13}$$

To implement this model, we first categorize our transaction costs into four categories based on the quartiles of their empirical distribution. The first category represents the lowest level of transaction costs (lower quartile) and the last category the highest level (upper quartile). Second, we select our variables through the forward method based on the p-value of the Wald chi-square test.[19] The reason for this choice is twofold. First, the materials to implement model selection methods are scarce in the ordinal logistic framework. Second, the intent here is to reduce the number of variables in order to avoid overfitting as well as having stable estimates (well-estimated conditional probabilities).[20]

Consistent with Kissell *et al.* (2004), who postulate that only trades with a certain level of contribution to the imbalance are costly to implement, we only consider split orders above the median of the contemporaneous contribution to the imbalance, defined as the volume of the split orders divided by the total market volume. For statistical robustness, we also consider only market members who placed at least 300[21] split orders in each duration group under investigation and in both directions. Finally, we remove duration and size from our set of explanatory variables and replace the Vol_1 variable with a dummy equal to 1 when the sequence of split orders contains volume from the first category of orders and 0 otherwise. Our motivation for these adjustments is that traders do not know ex ante the time and number of orders it will take to complete the trade, but they at least know whether they agree to be really aggressive in their execution strategy. All in all, we end up with a sample of 130,980 sequences of split orders from 10 market members.

[18] This procedure is accurate as long as collinearity among explanatory variables is not a concern. For each parameter, we computed the variance inflation factor (VIF), which measures the inflation in the variances of the parameter estimates caused by collinearity among the independent variables. Except for the trade imbalance VIF, which reaches 3.8, all the other VIF values are between 1 and 1.5, suggesting that multicollinearity is not a concern. The VIF option in the MODEL statement provides the VIFs. These factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the regressor (independent) variables

[19] The Wald statistic $\frac{\hat{\alpha}_j^2}{\sigma_{\alpha_j^2}}$ follows a χ^2 distribution with one degree of freedom, where $\sigma_{\alpha_j^2}$ is the variance of the coefficient α_j. The Wald chi-square test tests the null hypothesis that the estimate equals 0.

[20] There are, of course, other alternatives. For instance, Friedman *et al.* (2010) developed for the multinomial logistic regression a maximum likelihood estimation with a Lasso penalization algorithm available in the glmnet package of R software. One solution here would be not to enforce the ordinality and to use the multinomial framework. This remains correct as long as it is for forecasting purposes but does not make sense for the interpretation of the parameters.

[21] Moro *et al.* (2009) consider at least 1000 transactions per year. We consider at least 300 orders for 3 months.

Table III. Lasso selection regression results

	Panel A: Buy orders			Panel B: Sell orders		
	5 minutes	15 minutes	30 minutes	5 minutes	15 minutes	30 minutes
Order	47.4096	58.6161	70.7263	40.9121	51.2222	52.3857
Time	0.0037					
Vol_1	5E−06	7E−06		6E−06	4E−06	
Rspread	−22.8736	−19.5247	−19.3333	−23.3028	−19.5528	−22.9330
Hidden	−0.8989	−1.2309		−0.9217	−1.0675	
$\text{Buy}_{\text{cancel}}$	3.4443	4.7531	6.4865		−4.1224	−9.6495
$\text{Sell}_{\text{cancel}}$				3.1209	7.1352	
Opening	0.6834					
Oimb_1	0.6880	0.9368	1.4084	−0.7185	−1.0541	
Oimb_2		0.6756		−0.7547		
Oimb_3	0.2691					
Oimb_4	1.0900	2.3994	2.9756	−1.1027	−2.0502	−3.2045
Dimb_1	1.8084	2.2701	2.3430	−1.5352	−2.1554	−2.6256
Himb_1	1.3029					
Himb_2		1.3669		−0.8452		
Himb_3		1.7172				
Himb_4	0.9860	1.3690		−0.6481	−1.6505	

Note: This table reports the parameters of the transaction costs regression sample for buy and sell split orders consecutive to our adaptive Lasso selection. 'Order' refers to the square root of trade size relative to the average daily volume of the stock computed of 61 trading days. 'Time' is the time needed to complete the execution of the split orders. 'Vol_1' is the ratio of aggressive orders consuming the depth beyond the BOQ relative to the entire volume of the split orders. 'Rspread' is the relative spread prevailing at the beginning of the execution of each sequence of split orders. 'Hidden' is a dummy variable that takes the value of 1 if some hidden depth has been detected at the BOQ before the beginning of the sequence and 0 otherwise. '$\text{Buy}_{\text{cancel}}$' and '$\text{Sell}_{\text{cancel}}$' are the proportions of canceled orders on the bid and on the ask side, respectively. 'Opening' is a dummy that takes 1 if the sequence of split orders starts between the opening and 10:30 am and 0 otherwise. 'Oimb_j' are the order flow imbalances for the different i levels of aggressiveness computed during the 15-minute interval that precedes the start of the sequence of split orders. 'Dimb_1' is the scaled depth imbalance computed just before the start of the sequence of split orders. 'Himb_i' are the height imbalances computed at the i different levels of the limit order book just before the start of the sequences of split orders. All parameters are significant at the 1% level.

We use the first 2 months of data to estimate our model in-sample, while the last month is used for the out-of-sample estimation. The in-sample and out-of-sample periods clearly show different patterns. The market rose by around 4.8% during the in-sample period, whereas it decreased by almost 1% during the out-of-sample period.

EMPIRICAL WORK

Lasso regression results

Table III reports the OLS estimators of the best model identified through the Lasso selection procedure for each T3
panel (duration and direction). For comparison purposes, the results delivered by both the forward and the backward procedures based on the BIC criterion are provided in the Appendix, respectively in Tables A1, A2, A3 and A4.[22]

Firstly, with regard to our variables related to price movement, the aggressive order imbalance (Oimb_1) is relatively influential as it appears in five of the six regressions. As expected, it reveals a positive (negative) impact on transaction costs of buy (sell) split orders. In addition, less aggressive order imbalances (Oimb_2, Oimb_3) seem to be somewhat informative. Consistent with Tombeur and Wuyts (2011), we also observe a strong effect for the order flow imbalance of orders that improve the depth on the best quotes (Oimb_4).

The shape of the limit order book also explains part of the transaction costs. In particular, the height imbalance (Himb_i) is positively related to buy split orders transaction costs since an increase of its magnitude reflects a relatively more crowded order book on the bid side. The opposite holds for sell split orders. We should point out that the height imbalance at the first level is only significant for the sequences of buy split orders completed in less than 5 minutes. It reveals the difficulty of disentangling the 'price pressure' effect from the 'cost jumping' effect on the opposite side that we mentioned above. Indeed, an increase (decrease) of the height imbalance also comes from a less crowded

[22] The models are less parsimonious. They mainly include seasonality-related variables more often than the Lasso-selected model.

Table IV. Classification for the 5-minute split orders

5 minutes in-sample									
BUY					SELL				
N	Predicted				*N*	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	11,851	3,526	5,302	2,578	1	13,188	2,575	5,099	3,544
2	8,919	4,992	6,602	2,579	2	9,240	3,840	5,939	4,149
3	5,459	4,029	7,915	5,941	3	7,159	2,699	8,403	5,558
4	1,789	2,082	6,488	12,466	4	2,470	1,375	6,307	13,485
%	Predicted				%	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	51%	15%	23%	11%	1	54%	11%	21%	15%
2	39%	22%	29%	11%	2	40%	17%	26%	18%
3	23%	17%	34%	25%	3	30%	11%	35%	23%
4	8%	9%	28%	55%	4	10%	6%	27%	57%
5 minutes out-of-sample									
BUY					SELL				
N	Predicted				*N*	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	29,537	6,448	14,348	8,496	1	34,058	835	14,198	9,886
2	21,566	8,879	16,135	6,546	2	25,531	1,136	14,458	8,422
3	20,289	6,857	21,936	11,894	3	22,104	1,180	19,814	12,725
4	5,958	3,953	15,399	28,678	4	10,233	447	13,677	30,545
%	Predicted				%	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	50%	11%	24%	14%	1	58%	1%	24%	17%
2	41%	17%	30%	12%	2	52%	2%	29%	17%
3	33%	11%	36%	20%	3	40%	2%	35%	23%
4	11%	7%	29%	53%	4	19%	1%	25%	56%

Note: The above panel displays the classification results in-sample and out-of-sample for the ordered logit approach. It reports for both buy and sells split orders with a 5-minute duration the number and percentage of split orders with observed transaction costs in category i and predicted costs in category j $(i, j = 1, 2, 3, 4, 5)$. The results of each market member is weighted by the inverse of its percentage volume in the total sample in order to avoid any over-representation in the sample.

limit order book on the ask (bid) side and may increase the cost of a large order that walks down the order book. Moreover, orders that jump from the BOQ to the next one are very rare in our sample. Depth imbalance seems to play a less important role since the results are conclusive only for the depth imbalance at the best quotes (Dim_1). Besides, the impact could relate to the increase of liquidity or to any informational content as well. Indeed, an increase in the depth imbalance is the result of more depth on the bid side relative to the ask side. A sell (buy) order should be less (more) costly in such a situation.

In line with our expectations, transaction costs of buy (sell) orders are positively (negatively) related to the proportion of canceled orders on the bid side and negatively (positively) related to the proportion of canceled orders on the ask side. Consistent with the chasing hypothesis of Hasbrouck and Saar (2009), informed traders cancel their limit orders and switch to market orders. An increase in the cancellation rate on the bid side thus results in an increase in the aggressiveness on the same side, which triggers a reduction in transaction costs of buy split orders and an increase in transaction costs of sell split orders. Similarly, an increase in the cancellation rate on the ask side will trigger a reduction in transaction costs for buy split orders and the opposite for sell split orders. This result is consistent with the optimal execution strategies framework (Bertsimas and Lo, 1998; Almgren and Chriss, 1999, 2001; Krissell *et al.*, 2004) where traders seek to reach an optimum between transaction costs related to immediacy and the risk of being executed at a worse price. Indeed, traders trade more aggressively because they perceive a substantial increase in the risk of being executed at a less favorable price.

Secondly, the results for our liquidity determinants are mixed. Neither the average daily volume nor the market capitalization seem to play a role in explaining the transaction costs of split orders. The findings are similar for the

Table V. Classification for the 15-minute split orders

15 minutes in-sample									
BUY					SELL				
N	Predicted				*N*	Predicted			
Realized	1	2	3	4	Predicted	1	2	3	4
1	7,429	2,971	3,138	1,618	1	8,272	2,650	3,045	1,975
2	6,321	3,218	3,375	2,105	2	6,682	3,083	3,081	2,428
3	2,788	3,363	4,234	4,715	3	3,352	2,349	5,106	4,832
4	1,663	1,675	3,237	8,401	4	1,902	1,956	3,541	8,078
%	Predicted				%	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	49%	20%	21%	11%	1	52%	17%	19%	12%
2	42%	21%	22%	14%	2	44%	20%	20%	16%
3	18%	22%	28%	31%	3	21%	15%	33%	31%
4	11%	11%	22%	56%	4	12%	13%	23%	52%
15 minutes out-of-sample									
BUY					SELL				
N	Predicted				*N*	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	18,777	5,306	5,391	8,174	1	18,715	5,996	5,676	7,204
2	15,665	7,088	6,144	6,955	2	13,566	9,837	5,950	7,394
3	8,246	5,849	8,236	13,881	3	9,644	6,793	7,658	12,503
4	5,732	3,895	6,896	23,946	4	5,844	4,381	6,792	20,228
%	Predicted				%	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	50%	14%	14%	22%	1	50%	16%	15%	19%
2	44%	20%	17%	19%	2	37%	27%	16%	20%
3	23%	16%	23%	38%	3	26%	19%	21%	34%
4	14%	10%	17%	59%	4	16%	12%	18%	54%

Note: The above panel displays the classification results in-sample and out-of-sample for the ordered logit approach. It reports for both buy and sells split orders with a 15-minute duration the number and percentage of split orders with observed transaction costs in category i and predicted costs in category j $(i, j = 1, 2, 3, 4, 5)$. The results of each market member is weighted by the inverse of its percentage volume in the total sample in order to avoid any over-representation in the sample.

market place that does not even appear in the model. The opening variable is only significant for buy split orders executed at 5-minute intervals. We find no evidence of 'day-of-the week effect'.

Lastly, consistent with the literature, we observe a positive relationship between the magnitude of transaction costs and the order size (Order) of split orders. The degree of aggressiveness (Vol_1) is, unsurprisingly, positively related to transaction costs. Concerning the time to completion, we find mixed results as it appears significant only for the sequences of buy split orders whose duration does not exceed 5 minutes. We should point out that the number of orders to execute the total volume (N) does not significantly impact transaction costs of split orders. As for the detection of hidden liquidity, this plays a role for both buy and sell split orders up to 15 minutes. This effect is relatively long lived and confirms the findings of Detollenaere (2015). The relative spread is also significant, which is consistent with our expectations related to the mean reversion effect reported in Biais *et al.* (1995). Our volatility proxy is absent from the model, suggesting that it is not a key determinant of transaction costs for split orders.

Forecasting model

The classification of transaction costs delivered by our logit model is reported in Tables IV, V and VI, for the sequences of split orders depending on their duration. Results were computed for each market member separately and aggregated in a single output, wherein each market member's result is weighted by the inverse of its percentage volume in the total sample in order to avoid any over-representation bias.

Table VI. Classification for the 30-minute split orders

30 minutes in-sample

BUY *N*	Predicted				SELL *N*	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	5,073	1,801	1,830	1,108	1	5,098	1,752	1,421	1,320
2	3,570	2,905	1,526	1,794	2	3,387	2,244	2,200	1,623
3	1,899	2,166	3,017	2,700	3	1,927	1,862	3,094	2,653
4	1,008	1,188	2,004	5,509	4	1,057	951	1,855	2,653

%	Predicted				%	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	52%	18%	19%	11%	1	53%	18%	15%	14%
2	36%	30%	16%	18%	2	36%	24%	23%	17%
3	19%	22%	31%	28%	3	20%	20%	32%	28%
4	10%	12%	21%	57%	4	16%	15%	28%	41%

30 minutes out-of-sample

BUY *N*	Predicted				SELL *N*	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	8,861	5,153	3,820	3,049	1	12,664	3,757	4,461	3,303
2	5,620	8,872	4,625	3,828	2	9,114	5,308	5,051	3,917
3	3,908	5,545	5,839	6,476	3	5,493	3,312	7,912	6,271
4	2,363	4,265	4,378	11,408	4	2,791	2,760	5,623	10,284

%	Predicted				%	Predicted			
Realized	1	2	3	4	Realized	1	2	3	4
1	42%	25%	18%	15%	1	52%	16%	18%	14%
2	24%	39%	20%	17%	2	39%	23%	22%	17%
3	18%	25%	27%	30%	3	24%	14%	34%	27%
4	11%	19%	20%	51%	4	13%	13%	26%	48%

Note: The above panel displays the classification results in-sample and out-of-sample for the ordered logit approach. It reports for both buy and sell split orders with a 30-minute duration the number and percentage of split orders with observed transaction costs in category i and predicted costs in category j $(i, j = 1, 2, 3, 4, 5)$. The results of each market member is weighted by the inverse of its percentage volume in the total sample in order to avoid any over-representation in the sample.

Table VII. Quality measures for 5-minute split orders

5 minutes in-sample	Buy	Sell
Correctly classified	40%	41%
Low predicted high	17%	15%
High predicted low	21%	19%
>2 misclassified	21%	25%

5 minutes out-of-sample	Buy (vs. benchmark)	Sell (vs. benchmark)
Correctly classified	39 (32%)	39% (31%)
Low predicted high	15% (18%)	13% (18%)
High predicted low	19% (18%)	18% (18%)
>2 misclassified	26% (32%)	30% (33%)

Note: This table reports four quality measures of the model, which are computed for buy and sell split orders in-sample and out-of-sample, respectively, and out-of-sample for the linear discriminant analysis benchmark model. 'Correctly classified' is the percentage of split orders that are correctly classified. 'Low predicted high'/'High predicted low' refers to split orders that are classified above/below their level. '>2 misclassified' refers to the split orders that are misclassified with a difference of two levels. The results of each market member is weighted by the inverse of its percentage volume in the total sample in order to avoid any over-representation.

Table VIII. Quality measures for 15-minute split orders

	Buy	Sell
15 minutes in-sample		
Correctly classified	38.6%	39.37%
Low predicted high	18.4%	16.95%
High predicted low	21.4%	20.17%
>2 misclassified	21.6%	23.51%
15 minutes out-of-sample		
	Buy (vs. benchmark)	Sell (vs. benchmark)
Correctly classified	39% (32%)	38% (31%)
Low predicted high	17% (22%)	16% (21%)
High predicted low	19% (17%)	18% (16%)
>2 misclassified	26% (30%)	27% (31%)

Note: This table reports four quality measures of the model, which are computed for buy and sell split orders in-sample and out-of-sample, respectively, and out-of-sample for the linear discriminant analysis benchmark model. 'Correctly classified' is the percentage of split orders that are correctly classified. 'Low predicted high'/'High predicted low' refers to split orders that are classified above/below their level. '>2 misclassified' refers to the split orders that are misclassified with a difference of two levels. The results of each market member is weighted by the inverse of its percentage volume in the total sample in order to avoid any over-representation.

Table IX. Quality measures for 30-minute split orders

	Buy	Sell
30 minutes in sample		
Correctly classified	42%	37%
Low predicted high	15%	19%
High predicted low	20%	20%
>2 misclassified	23%	24%
30 minutes out of sample		
	Buy (vs. benchmark)	Sell (vs. benchmark)
Correctly classified	40% (30%)	39% (29%)
Low predicted high	18% (21%)	16% (23%)
High predicted low	18% (17%)	20%(17%)
>2 misclassified	24% (33%)	25% (31%)

Note: This table reports four quality measures of the model, which are computed for buy and sell split orders in and out-of-sample, respectively, and out-of-sample for the linear discriminant analysis benchmark model. 'Correctly classified' is the percentage of split orders that are correctly classified. 'Low predicted high'/'High predicted low' refers to split orders that are classified above/below their level. '>2 misclassified' refers to the split orders that are misclassified with a difference of two levels. The results of each market member are weighted by the inverse of its percentage volume in the total sample in order to avoid any over-representation.

Whatever the time to execution and the order direction, our logit model performs very well, both in-sample and out-of-sample. In particular, we are able to predict out-of-sample from 48% (for sell orders in Table VI) to 59% (for buy orders in Table V) of the most expensive trades. Similarly, we correctly classify around 50% of the cheapest trades up to 30 minutes out-of-sample. If we compare the model performance to the performance of a naive model assigning a trade to a category $i = 1, \ldots, 4$ with probability 1/4 (25%), we note that we strongly outperform the naive model in almost all cases. In fact, we only miss the naive model 25% target by a few percent for categories 2 and 3 in some rare cases.

Tables VII, VIII and IX exhibit corresponding quality measures that attest to the good performance of our model, both in-sample and out-of-sample. Specifically, they show that we reach an average percentage of correctly classified orders of about 40%, as well as a relatively low percentage of serious misclassification of about 25%. For comparison purposes, we also report the same quality measures out-of-sample for a benchmark model. This model is a linear discriminant analysis that relies on standard transaction cost determinants,[23] highlighted in Chan and Lakonishok (1995). The comparison of our model with our benchmark confirms the pertinence of our approach.

[23] Consistent with Chan and Lakonishok (1995), transaction costs are explained by the split order size in relation to the normal daily volume that captures the trade difficulty, the market capitalization and the identity of the trader.

CONCLUSION

This paper provides an analysis of the impact of microstructure variables on the transaction costs for split orders on 171 Euronext large cap stocks.

First, using the adaptive Lasso selection of Zou (2006), we identify the set of variables that best explain transaction costs of split orders. The results suggest that the order flow imbalance and the proportion of fleeting orders are significant determinants. The shape of the limit order book seems also to impact transaction costs but to a lesser extent. The size of the sequence of split orders as well as its degree of aggressiveness are positively related to transaction costs. As for the detection of hidden liquidity, this plays a role for both buy and sell split orders up to 15 minutes.

Second, in order to demonstrate the forecasting power of microstructure variables, we propose an alternative ordinal logistic model that aims at classifying ex ante split orders into 'buckets' that represent different levels of transaction costs. We then use this model to check whether investors could effectively predict the magnitude of their transaction costs based on microstructure information. The results are conclusive both in-sample and out-of-sample. In particular, we are able to predict on average more than 50% of the most expensive trades. We reach an average percentage of correctly classified orders of about 40% as well as a relatively low percentage of serious misclassification of about 25%.

All these findings provide evidence that microstructure variables are of great importance for managing the implicit transaction costs of split orders.

ACKNOWLEDGEMENT

The authors are grateful to Euronext in Paris for providing the data. They thank Carole Gresse, Jean-Christophe Meyfredi, Mikael Petitjean as well as anonymous referees for providing useful comments and suggestions. Any remaining errors are the responsibility of the authors.

REFERENCES

Abhyankar A, Ghosh D, Levin E, Limmack R. 2003. Bid–ask spreads, trading volume and volatility: intra-day evidence from the London Stock Exchange. *Journal of Business Finance and Accounting* **24**(3): 343–362.

Almgren R, Chriss N. 1999. Value under liquidation. *Risk* **12**(12): 61–63.

Almgren R, Chriss N. 2001. Optimal execution of portfolio transactions. *Journal of Risk* **3**: 5–40.

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

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

Bikker J, Spierdijk L, Van Der Sluis P. 2007. Market impact costs of institutional equity trades. *Journal of International Money and Finance* **26**(6): 974–1000.

Bikker J, Spierdijk L, Hoevenaars R, Van der Sluis P. 2008. Forecasting market impact costs and identifying expensive trades. *Journal of Forecasting* **27**(1): 21–39.

Cao C, Hansch O, Wang X. 2009. The information content of an open limit-order book. *Journal of Futures Markets* **29**(1): 16–41.

Chan K, Christie W, Schultz P. 1995. Market structure and the intraday pattern of bid–ask spreads for NASDAq securities. *Journal of Business* **68**(1): 35–60.

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

Chan L, Lakonishok J. 1997. Institutional equity trading costs: NYSE versus NASDAQ. *Journal of Finance* **52**(2): 713–735.

Chordia T, Subrahmanyam A. 2004. Order imbalance and individual stock returns: theory and evidence. *Journal of Financial Economics* **72**(3): 485–518.

Chordia T, Roll R, Subrahmanyam A. 2002. Order imbalance, liquidity, and market returns. *Journal of Financial Economics* **65**(1): 111–130.

De Winne R, D'Hondt C. 2007. Hide-and-seek in the market: placing and detecting hidden orders. *Review of Finance* **11**(4): 663–692.

Detollenaere B. 2015. *The virtue of impatience when trading split orders against hidden liquidity*, Working paper.

Domowitz I, Glen J, Madhavan A. 2001. Liquidity, volatility and equity trading costs across countries and over time. *International Finance* **4**(2): 221–255.

Donoho DL. 2006. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics* **59**(6): 797–829.

Donoho DL, Elad M. 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences USA* **100**(5): 2197–2202.

Donoho DL, Johnstone IM, Kerkyacharian G, Picard D. 1995. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society, Series B* **57**(2): 301–369.

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *Annals of Statistics* **32**(2): 407–499.

Foster F, Viswanathan S. 1990. A theory of the interday variations in volume, variance, and trading costs in securities markets. *Review of Financial Studies* **3**(4): 593–624.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1–22.

Gomber P, Schweickert U, Theissen E. 2004. Zooming in on liquidity. *EFA 2004 Maastricht Meetings Paper No.1805*.

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

Hasbrouck J, Saar G. 2009. Technology and liquidity provision: the blurring of traditional definitions. *Journal of Financial Markets* **12**(2): 143–172.

Keim D, Madhavan A. 1997. Transactions costs and investment style: an inter-exchange analysis of institutional equity trades. *Journal of Financial Economics* **46**(3): 265–292.

Kissell R, Glantz M, Malamut R. 2003. *Optimal Trading Strategies: Quantitative Approaches for Managing Market Impact and Trading Risk*. Amacom: New York.
Kissell R, Glantz M, Malamut R. 2004. A practical framework for estimating transaction costs and developing optimal trading strategies to achieve best execution. *Finance Research Letters* **1**(1): 35–46.
Lee Y, Liu Y, Roll R, Subrahmanyam A. 2004. Order imbalances and market efficiency: evidence from the Taiwan Stock Exchange. *Journal of Financial and Quantitative Analysis* **39**(2): 327–341.
Lillo F, Moro E, Vaglica G, Mantegna R. 2008. Specialization and herding behavior of trading firms in a financial market. *New Journal of Physics* **10**(4): 043019.
Markowitz H. 1952. Portfolio selection. *Journal of Finance* **7**(1): 77–91.
Meinshausen N, Bühlmann P. 2006. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34**(3): 1436–1462.
Moro E, Vicente J, Moyano LG, Gerig A, Farmer JD, Vaglica G, Lillo F, Mantegna RN. 2009. Market impact and trading profile of hidden orders in stock markets. *Physical Review E* **80**(6): 066102.
Nielsson U. 2009. Stock exchange merger and liquidity: the case of Euronext. *Journal of Financial Markets* **12**(2): 229–267.
O'Hara M. 2015. High frequency market microstructure. *Journal of Financial Economics* **116**(2): 257–270.
Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**: 267–288.
Tombeur G, Wuyts G. 2011. *Does information in the limit order book help to predict returns?* Working paper presented at Erasmus Liquidity Conference (Doctoral Symposium- Rotterdam, The Netherlands, 30 June-1 July 2011).
Vaglica G, Lillo F, Moro E, Mantegna R. 2008. Scaling laws of strategic behavior and size heterogeneity in agent dynamics. *Physical Review E* **77**(3): 036110.
Zou H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476): 1418–1429.

APPENDIX

Table A1. Forward selection regression results: buy split orders

	5 minutes		15 minutes		30 minutes	
	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.
Order	46.3523	***	59.14576	***	72.52115	***
N	0.04765	***				
Time	0.00354	***				
Vol_1	4.93E−06	***	6.69E−06	***		
Rspread	−22.7311	***	−19.698	***	−18.3346	***
Hidden	−0.9609	***	−1.23027	***	−1.33378	***
HL	−0.28559	***			−1.92748	***
$\mathrm{Buy}_{\mathrm{cancel}}$	3.15694	***	5.01215	***	6.30289	***
$\mathrm{Sell}_{\mathrm{cancel}}$					−4.04442	***
Opening	0.72656	***			1.03069	***
Closing			−0.52716	***		
PT	2.35202	***				
Oimb_1	0.63943	***	0.9273	***	1.41047	***
Oimb_2	0.37665	***	0.67111	***		
Oimb_3	0.25022	***			0.60662	***
Oimb_4	1.1107	***	2.39988	***	2.47697	***
Dimb_1	1.80062	***	2.24209	***	2.36316	***
Dimb_2	0.42314	***	0.53885	***		
Dimb_5	0.21652	***				
Himb_2	1.22083	***	1.37699	***	1.62022	***
Himb_3	0.57478	***	1.71063	***	1.41487	***
Himb_4	0.57752	***	1.36362	***	1.58883	***
Himb_5	0.81985	***			1.95265	***
Tue					−1.87473	***
Fri					−0.90372	***

Note: This table reports the parameters of the transaction costs regression sample for buy split orders consecutive to our forward selection. 'Order' refers to the square root of trade size relative to the average daily volume of the stock computed of 61 trading days. 'N' refers to the number of split orders. 'Time' is the time needed to complete the execution of the split orders. 'Vol_1' is the ratio of aggressive orders consuming the depth beyond the BOQ relative to the entire volume of the split orders. 'Rspread' is the relative spread prevailing at the beginning of the execution of each sequence of split orders. 'Hidden' is a dummy variable that takes the value of one if some hidden depth has been detected at the BOQ before the beginning of the sequence and zero otherwise. 'HL' is the high–low of transaction prices during the 15-minute interval that precedes the split orders scaled by the last transaction price of the interval. $\mathrm{Buy}_{\mathrm{cancel}}$ and $\mathrm{Sell}_{\mathrm{cancel}}$ are the proportion of canceled orders on the bid and on the ask side, respectively. 'Opening' and 'Closing' are dummies that control for the intraday seasonality. 'PT' is a dummy that controls for the marketplace. 'Oimb_j' are the order flow imbalances for the different j levels of aggressiveness computed during the 15-minute interval that precedes the start of the sequence of split orders. 'Dimb_1' is the scaled depth imbalance computed just before the start of the sequence of split orders. 'Himb_i' are the heights imbalance computed at the i different levels of the limit order book just before the start of the sequences of split orders. 'Tue'/'Fri' is a dummy that takes the value of one if the transaction takes place on Tuesday/Friday and zero otherwise. *** indicates parameters significant at the 1% level.

Table A2. Forward selection regression results: sell split orders

	5 minutes		15 minutes		30 minutes	
	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.
Order	38.18145	***	52.22541	***	45.55553	***
N	0.08963	***				
Time	0.0024	***				
Vol_1	0.00000559	***	4.43E−06	***	6.1E−06	***
Rspread	−23.5886	***	−20.0327	***	−23.8795	***
Hidden	−0.95184	***	−1.15392	***	−1.45633	***
HL	−0.27824	***			−1.37944	***
$\text{Buy}_{\text{cancel}}$			−2.93492	***	−7.70928	***
$\text{Sell}_{\text{cancel}}$	3.39553		6.839	***	5.6977	***
Opening	0.50454	***				
Closing	−0.31553	***	−0.77264	***	−1.34009	***
Afternoon					−1.53634	***
PT	3.3236	***				
FR			0.43238	***		
Oimb_1	−0.69851	***	−0.97571	***	−1.22332	***
Oimb_2	−0.62234	***	−0.67503	***	−1.50398	***
Oimb_3	−0.18593	***	−0.36765	***		
Oimb_4	−1.01127	***	−2.02097	***	−2.44321	***
Dimb_1	−1.52349	***	−2.14366	***	−2.5393	***
Dimb_2	−0.39387	***	−0.84553	***	−0.70385	***
Himb_2	−0.67323	***	−1.01971	***		
Himb_3	−0.82344	***	−1.24324	***	−1.80164	***
Himb_4	−0.6558	***	−1.30647	***	−1.95813	***
Himb_5	−0.63642	***	−1.56628	***	−2.17087	***
Timb	−0.10909				0.18986	
Tue			0.49317	***	1.02429	***
Logcapi					−0.37983	***

Note: This table reports the parameters of the transaction costs regression sample for sell split orders consecutive to our forward selection. 'Order' refers to the square root of trade size relative to the average daily volume of the stock computed of 61 trading days. 'N' refers to the number of split orders. 'Time' is the time needed to complete the execution of the split orders. Vol_1 is the ratio of aggressive orders consuming the depth beyond the BOQ relative to the entire volume of the split orders. 'Rspread' is the relative spread prevailing at the beginning of the execution of each sequence of split orders. 'Hidden' is a dummy variable that takes the value of one if some hidden depth has been detected at the BOQ before the beginning of the sequence and zero otherwise. 'HL' is the high–low of transaction prices during the 15-minute interval that precedes the split orders scaled by the last transaction price of the interval. '$\text{Buy}_{\text{cancel}}$' and '$\text{Sell}_{\text{cancel}}$' are the proportion of canceled orders on the bid and on the ask side, respectively. 'Opening', 'Afternoon' and 'Closing' are dummies that control for the intraday seasonality. 'PT' and 'FR' are dummies that control for the market place. 'Oimb_j' are the order flow imbalances for the different j levels of aggressiveness computed during the 15-minute interval that precedes the start of the sequence of split orders. 'Dimb_1' is the scaled depth imbalance computed just before the start of the sequence of split orders. 'Himb_i' are the heights imbalance computed at the i different levels of the limit order book just before the start of the sequences of split orders. 'Timb' stands for the trade imbalance during the 15-minute interval preceding the inception of the sequence of split orders. 'Tue' is a dummy that takes the value of one if the transaction takes place on Tuesday and zero otherwise. 'Logcapi' is the logarithm of the market capitalization in EUR. *** indicates parameters significant at the 1% level.

Table A3. Backward selection regression results: buy split orders

	5 minutes		15 minutes		30 minutes	
	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.
Order	46.3523	***	59.14576	***	72.52115	***
N	0.04765	***				
Time	0.00354	***				
Vol_1	0.00000493	***	6.69E−06	***		
Rspread	−22.73109	***	−19.698	***	−18.3346	***
Hidden	−0.9609	***	−1.23027	***	−1.33378	***
HL	−0.28559	***			−1.92748	***
$\mathrm{Buy}_{\mathrm{cancel}}$	3.15694	***	5.01215	***	6.30289	***
$\mathrm{Sell}_{\mathrm{cancel}}$					−4.04442	***
Opening	0.72656	***			1.03069	***
Closing			−0.52716	***		
PT	2.35202	***				
FR						
Oimb_1	0.63943	***	0.9273	***	1.41047	***
Oimb_2	0.37665	***	0.67111	***		
Oimb_3	0.25022	***			0.60662	***
Oimb_4	1.1107	***	2.39988	***	2.47697	***
Dimb_1	1.80062	***	2.24209	***	2.36316	***
Dimb_2	0.42314	***	0.53885	***		
Dimb_5	0.21652	***				
Himb_2	1.22083	***			1.62022	***
Himb_3	0.57478	***	1.37699	***	1.41487	***
Himb_4	0.57752	***	1.71063	***	1.58883	***
Himb_5	0.81985	***	1.36362	***	1.95265	***
Tue	−1.87473	***				
Fri	−0.90372	***				

Note: This table reports the parameters of the transaction costs regression sample for buy split orders consecutive to our backward selection. 'Order' refers to the square root of trade size relative to the average daily volume of the stock computed of 61 trading days. 'N' refers to the number of split orders. 'Time' is the time needed to complete the execution of the split orders. 'Vol_1' is the ratio of aggressive orders consuming the depth beyond the BOQ relative to the entire volume of the split orders. 'Rspread' is the relative spread prevailing at the beginning of the execution of each sequence of split orders. 'Hidden' is a dummy variable that takes the value of one if some hidden depth has been detected at the BOQ before the beginning of the sequence and zero otherwise. 'HL' is the high–low of transaction prices during the 15-minute interval that precedes the split orders scaled by the last transaction price of the interval. '$\mathrm{Buy}_{\mathrm{cancel}}$' and '$\mathrm{Sell}_{\mathrm{cancel}}$' are the proportion of canceled orders on the bid and on the ask side, respectively. 'Opening' and 'Closing' are dummies that control for the intraday seasonality. 'PT' and 'FR' are dummies that control for the market place. 'Oimb_j' are the order flow imbalances for the different j levels of aggressiveness computed during the 15-minute interval that precedes the start of the sequence of split orders. 'Dimb_1' is the scaled depth imbalance computed just before the start of the sequence of split orders. 'Himb_i' are the heights imbalance computed at the i different levels of the limit order book just before the start of the sequences of split orders. 'Tue'/'Fri' is a dummy that takes the value of one if the transaction takes place on Tuesday/Friday and zero otherwise. *** indicates parameters significant at the 1% level.

Table A4. Backward selection regression results: sell split orders

	5 minutes		15 minutes		30 minutes	
	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.
Order	38.19965	***	52.22541		45.52848	***
N	0.0896	***				
Time	0.0024	***				
Vol_1	0.00000559	***	4.43E−06	***	6.1E−06	***
Rspread	−23.58212	***	−20.0327	***	−23.8858	***
Hidden	−0.94975	***	−1.15392	***	−1.45968	***
HL	−0.28023	***			−1.3746	***
Buy_{cancel}			−2.93492	***	−7.69885	***
$Sell_{cancel}$	3.40396	***	6.839	***	5.67819	***
Opening	0.50496	***				
Closing	−0.31603	***	−0.77264	***	−1.33896	***
Afternoon					−1.53572	***
PT	3.3344	***				
FR			0.43238	***		
$Oimb_1$	−0.73037	***	−0.97571	***	−1.16853	***
$Oimb_2$	−0.6862	***	−0.67503	***	−1.39036	***
$Oimb_3$	−0.18829	***	−0.36765	***		
$Oimb_4$	−1.01793	***	−2.02097	***	−2.43242	***
$Dimb_1$	−1.52377	***	−2.14366	***	−2.53861	***
$Dimb_2$	−0.3949	***	−0.84553	***	−0.70289	***
$Himb_2$	−0.67254	***	−1.01971	***		
$Himb_3$	−0.82366	***	−1.24324	***	−1.79986	***
$Himb_4$	−0.65576	***	−1.30647	***	−1.95836	***
$Himb_5$	−0.63655	***	−1.56628	***	−2.17183	***
Tue			0.49317	***	1.0245	***
Logcapi					−0.38002	***

Note: This table reports the parameters of the transaction costs regression sample for sell split orders consecutive to our backward selection. 'Order' refers to the square root of trade size relative to the average daily volume of the stock computed of 61 trading days. 'N' refers to the number of split orders. 'Time' is the time needed to complete the execution of the split orders. 'Vol_1' is the ratio of aggressive orders consuming the depth beyond the BOQ relative to the entire volume of the split orders. 'Rspread' is the relative spread prevailing at the beginning of the execution of each sequence of split orders. 'Hidden' is a dummy variable that takes the value of one if some hidden depth has been detected at the BOQ before the beginning of the sequence and zero otherwise. 'HL' is the high–low of transaction prices during the 15-minute interval that precedes the split orders scaled by the last transaction price of the interval. 'Buy_{cancel}' and '$Sell_{cancel}$' are the proportion of canceled orders on the bid and on the ask side, respectively. 'Opening', 'Afternoon' and 'Closing' are dummies that control for the intraday seasonality. 'PT' and 'FR' are dummies that control for the market place. '$Oimb_j$' are the order flow imbalances for the different j levels of aggressiveness computed during the 15-minute interval that precedes the start of the sequence of split orders. '$Dimb_1$' is the scaled depth imbalance computed just before the start of the sequence of split orders. '$Himb_i$' are the heights imbalance computed at the i different levels of the limit order book just before the start of the sequences of split orders. 'Tue' is a dummy that takes the value of one if the transaction takes place on Tuesday and zero otherwise. 'Logcapi' is the logarithm of the market capitalization in EUR. *** indicates parameters significant at the 1% level.

Authors' biographies:
Benoit Detollenaere is a former PhD student at Louvain School of Management – Université catholique de Louvain. This paper is part of his PhD thesis.

Catherine D'Hondt is a Professor of finance at Louvain School of Management – Université catholique de Louvain.

Authors' addresses:
Benoit Detollenaere and **Catherine D'Hondt**, Louvain School of Management - Université catholique de Louvain, 151 Chaussée de Binche, 7000 Mons, Belgium.

