Assessing accuracy of trade side classification rules. Methods, data, and problems

Joanna Olbryś[1], Michał Mursztyn[2]

Abstract

Trade side classification algorithms enable us to assign the side that initiates a transaction and distinguish between the so-called buyer- and seller-initiated trades. According to the literature, such classification is essential to assess both market liquidity and dimensions of market liquidity based on high frequency intraday data. The main problem is that trade and quote data is not publicly available for many stock markets and researchers have to utilize indirect methods to infer trade side. The aim of this paper is to investigate major problems in assessing accuracy of trade side classification algorithms. We evaluate and compare four most frequently utilized procedures using intraday data for 105 companies from the Warsaw Stock Exchange (WSE). Moreover, an analysis of the robustness of the results is provided over the whole sample period from January 2, 2005 to December 30, 2016, and three consecutive sub-periods of equal size, covering the pre-crisis, crisis, and post-crisis periods. The empirical experiment shows that the Lee-Ready (1991) algorithm and tick rule perform better than other methods on the WSE, regardless of the choice of the sample.

Keywords: *market microstructure, high frequency data, trades, quotes, trade side classification procedures*

JEL Classification:C10, C55, C63, C81, G1

DOI: 10.14659/SEMF.2018.01.34

1 Introduction

High frequency data is important in studying a variety of issues related to trading processes and market microstructure. To measure both market liquidity and dimensions of market liquidity based on intraday data, it is essential to recognize the side that initiates a transaction, and to distinguish between the so-called buyer- and seller-initiated trades. In the absence of information regarding whether a trade is buyer or seller initiated, many researchers employ various classification procedures to infer trade side on financial markets in the world.

On the contrary to research conducted on international stock markets, there are only a few studies that make use of trade side classification algorithms for the WSE data. For instance, Nowak (2017) analyses the problem of asset pricing on the basis of high frequency data for the WSE. She uses the quote rule to classify the side that initiates a transaction. Olbryś (2017) conducts the study of interaction between market depth and market tightness on the WSE and

[1] Corresponding author: Bialystok University of Technology, Faculty of Computer Science, Wiejska 45 A, 15-351 Białystok, j.olbrys@pb.edu.pl.

[2]Bialystok University of Technology, Faculty of Computer Science, Wiejska 45 A, 15-351 Białystok, m.mursztyn@student.pb.edu.pl.

she uses the Lee and Ready (1991)algorithm to infer trade sides. Olbryś and Mursztyn (2017) estimate and analyse selected liquidity proxies derived from intraday data and supported by the Lee-Ready algorithm inferring the initiator of a trade.

As the subject of accuracy of a trade side classification is crucial for market microstructure research, the goal of this paper is to investigate major problems in assessing accuracy of selected algorithms inferring the initiator of a trade. We evaluate and compare four most frequently utilized procedures based on intraday data for 105 WSE-listed companies in the period from January 2, 2005 to December 30, 2016. Moreover, an analysis of the robustness of the obtained results is provided over the whole sample period and three consecutive sub-samples of equal size, namely the pre-crisis, crisis, and post-crisis periods. The empirical experiment shows that the Lee-Ready algorithm and the tick rule perform better than other methods on the WSE, regardless of the choice of the sample.

The remainder of the study is organized as follows. Section 2 specifies algorithms used in the research and presents a brief literature review in the context of assessing accuracy of trade side classification procedures. Section 3 describes the data. In Section 4, we present and discuss the empirical experiment for high frequency data from the WSE. The last section encompasses the conducted research with a brief summary.

2 Trade side classification procedures

The goal of trade side classification is to determine the initiator of a transaction and to classify trades as being either buyer or seller motivated. However, a formal definition of a trade initiator is rarely stated in the literature. Moreover, researchers use various definitions of a trade initiator based presumably on data availability. For example, the so-called 'immediacy' definition describes initiators as traders who demand immediate execution (Lee and Radhakrishna, 2000). Thus, it is usually possible to classify each trade as either buyer- or seller-initiated. Odders-White (2000) considers the last arriving order to be trade initiator (the so-called ‘chronological’ definition). These two definitions are equivalent in many cases. In both of them, the initiator is the person who caused the transaction to occur. Theissen (2001) proposes somewhat different definition based on a position taken by a specialist, which is appropriate for a hybrid equity market.

Table 1. Selected trade side classification procedures.

Rule	Conditions	
QR	Trade is classified as buyer-initiated if $P_t > P_t^{mid}$	Trade is classified as seller-initiated if $P_t < P_t^{mid}$
	If $P_t = P_t^{mid}$ then a trade is not classified.	
TR	Trade is classified as buyer-initiated if $P_t > P_{t-1}$	Trade is classified as seller-initiated if $P_t < P_{t-1}$
	When $P_t = P_{t-1}$, a trade is signed using the previous transaction price. If the sign of the last non-zero price change is positive (negative) then the trade is classified as buyer-initiated (seller-initiated).	
	I stage	
LR	Trade is classified as buyer-initiated if $P_t > P_t^{mid}$	Trade is classified as seller-initiated if $P_t < P_t^{mid}$
	If $P_t = P_t^{mid}$ then:	
	II stage	
	Trade is classified as buyer-initiated if $P_t^{mid} > P_{t-1}$	Trade is classified as seller-initiated if $P_t^{mid} < P_{t-1}$
	When $P_t^{mid} = P_{t-1}$, the decision is made using the sign of the last non-zero price change. If $P_t > P_{t-k}$ then it is buyer-initiated. If $P_t < P_{t-k}$ then it is seller-initiated.	
EMO	I stage	
	Trade is classified as buyer-initiated if $P_t = P_t(a) = P_t^L$	Trade is classified as seller-initiated if $P_t = P_t(b) = P_t^H$
	II stage	
	Trades with prices different from best ask and best bid prices are categorized by the first stage of the tick rule and P_t is compared to P_{t-1}. If $P_t > P_{t-1}$ then trade is classified as buyer-initiated. If $P_t < P_{t-1}$ then trade is classified as seller-initiated.	

Notes: P_t - the transaction price at the time *t*, approximated by the closing price.

$P_t(a) = P_t^L$ - the best ask price, approximated by the lowest price at the time *t*.

$P_t(b) = P_t^H$ - the best bid price, approximated by the highest price at the time *t*.

$P_t^{mid} = \frac{P_t(a)+P_t(b)}{2} = \frac{P_t^L+P_t^H}{2}$ - the so-called quoted midpoint at the time *t*.

Table 2. Assessing accuracy of selected trade side classification procedures.

Market	Procedures	Data	Results	Source
Australian Stock Exchange (ASX)	TR	SEATS[3] database	An accuracy of ≈74% (TR)	Aitken and Frino (1996)
NASDAQ	QR, TR, LR, EMO	TAQ[4] database	An accuracy of 76.4% (QR), 77.66% (TR), 81.05% (LR),81.9% (EMO)	Ellis et al. (2000)
NYSE	QR, TR, LR	TORQ[5]database	An accuracy of ≈75% (QR), ≈79% (TR), ≈85% (LR)	Odders-White (2000)
NYSE	TR, LR	TORQ database	An accuracy between 83% and 84% (both TR and LR)	Finucane (2000)
Frankfurt Stock Exchange (FSE)	TR, LR	Handels-überwachungsstelle	An accuracy of 72.8% (LR) and 72.2% (TR)	Theissen (2001)
Taiwan Stock Exchange (TWSE)	QR, TR, LR, EMO, and others	TWSE database	An accuracy of 92.75% (QR), 74.18% (TR), 96.46% (LR), 94.97% (EMO)	Lu and Wei (2009)
NYSE, NASDAQ	QR, TR, LR	TAQ, CRSP[6] databases	QR, TR, and LR classify a majority of short sales as buyer-initiated	Asquith et al. (2010)
NASDAQ	TR, LR, BVC	ITCH[7], TAQ databases	An accuracy of 90.8% (TR), 92.6% (LR), 79.7% (BVC)	Chakrabarty et al. (2015)

There are quite many trade side classification methods described in the literature, but the most frequently used are: the quote rule (QR), the tick rule (TR), the Lee-Ready (LR) algorithm, and the Ellis-Michaely-O'Hara (EMO) algorithm (see Table 1.)[8]. Easley et al.

[3]SEATS – the Stock Exchange Automated Trading System , the Australian Stock Exchange.
[4]TAQ –the Trades and Quotes database.
[5]TORQ–the Trades, Orders, Reports, and Quotes database.
[6] CRSP –the Center for Research in Security Prices.
[7]ITCH – the ITCH database provided by NASDAQ.
[8]The implementation of four trade side classification algorithms is presented in detail in the paper (Olbryś and Mursztyn, 2015).

(2013) propose the BVC (Bulk Volume Classification) procedure, but this approach apportions trades into buy volume and sell volume, and does not assign trades to be either buyer- or seller-initiated.

The accuracy of trade classification algorithms has been examined in many studies, however the results have been diverse for various stock markets. Table 2. summarizes the empirical results of assessing the accuracy of selected trade side classification procedures on some international stock markets. We focus on the QR, TR, LR, and EMO methods.

As presented in Table 2., there are some discrepancies in trade side classification results between markets. The possible explanation of them is that stock market structure, institutional differences between markets, and trading mechanisms may affect the accuracy of classification procedures. Specifically, the U.S. stock markets (NYSE, AMEX and NASDAQ) and the FSE are hybrid and primarily quote-driven markets in which market-makers/specialists play a prominent role, while the ASX, TWSE and WSE are order-driven markets with fully automated trading systems.

3 Data description

In this study, high frequency data ‘rounded to the nearest second’ from the WSE (available at www.bossa.pl) for the 105 WSE-listed companies was utilized. The dataset contains the opening, high, low and closing prices, as well as volume for a security over one unit of time. The whole sample covers the period from January 2, 2005 to December 30, 2016 (3005 trading days). To verify the robustness of the obtained results, the comparison of trade side classification rules is provided both for the whole sample and over three consecutive subsamples, each of equal size (436 trading days) (see Olbryś and Mursztyn, 2015):

1. The pre-crisis period from September 6, 2005 to May 31, 2007.
2. The crisis period from June 1, 2007 to February 27, 2009.
3. The post-crisis period from March 2, 2009 to November 19, 2010.

The crisis period on the WSE was formally defined based on the paper (Olbryś and Majewska, 2015), in which the statistical method for the formal identification of market states was employed.

When forming the database, we included only the securities that were listed on the WSE for the whole sample period since December 31, 2004, and were not suspended. The 139 WSE companies met these basic conditions, and they were initially selected. Next, to mitigate the non-trading problem on the WSE, we excluded the stocks that exhibited extraordinarily many non-traded days during the whole sample period, precisely, above 300 zeros in daily volume, which constituted about 10% of all 3005 trading days. The database consists of 105 companies regarding the presented way of their selection. The dataset is large. Within the trading days during the whole sample period, the total number of records in the database is equal to 35 307 993.

4 The empirical experiment on the Warsaw Stock Exchange

We evaluate and compare four most frequently utilized trade side classification rules for the data from the WSE. An analysis of the robustness of the results is provided over the whole sample period and three adjacent sub-periods, each of equal size. The average percentage values of classified and unclassified trades in the case of all trade classification procedures, for the whole group of 105 WSE-listed stocks are presented in Table 3. The empirical findings indicate that usefulness of various trade side classification methods on the WSE is not qualitatively the same, whereas the results turn out to be robust to the choice of the period.

Specifically, the tick rule (TR) and the Lee-Ready (LR) algorithm are more appropriate compared to the quote rule (QR) and EMO procedure. In the case of the TR and LR methods, the percentage of unclassified transactions is relatively low and similar, which is consistent with the literature. For example, Theissen (2001) points out that the Lee-Ready method classifies transactions quite correctly, but the simpler tick test performs almost equally well. The amount of buyer- and seller-initiated trades is almost equal, with a little predominance of buyer-initiated in all investigated periods. This evidence is in accordance with the literature, as it is demonstrated in some papers that short sales are sometimes misclassified as buyer-initiated by some trade side classification algorithms (Asquith, 2010).

On the contrary, applicability and accuracy of the QR and the EMO procedures is rather low, with high percentage of unclassified trades for all companies, regardless of a firm’s size and the choice of the period. In our opinion, this phenomenon on the WSE is caused by the problem of relatively many trades for which high and low prices are equal over one unit of time. Moreover, the EMO method was proposed for the NASDAQ, which is a hybrid market, while the WSE is a pure order-driven market. Hence, the WSE differs from the NYSE and the NASDAQ, and therefore the empirical results based on the U.S. stock markets are not comparable to Polish stock market.

Table 3. The average percentage values of classified and unclassified trades for the whole group of the 105 WSE-listed companies (the LR, TR, QR, and EMO procedures). The best results for the LR algorithm are marked in bold.

Rule	Period	Mean value of buyer-initiated trades (%)	Mean value of seller-initiated trades (%)	Mean value of unclassified trades (%)
LR	Whole sample	48.06%	45.58%	**6.36%**
	Pre-crisis	49.08%	45.41%	**5.51%**
	Crisis	46.47%	47.03%	**6.50%**
	Post-crisis	47.77%	44.85%	**7.38%**
TR	Whole sample	47.95%	45.60%	6.45%
	Pre-crisis	49.07%	45.35%	5.58%
	Crisis	46.28%	47.11%	6.61%
	Post-crisis	47.64%	44.86%	7.50%
QR	Whole sample	5.88%	6.35%	87.77%
	Pre-crisis	5.95%	6.44%	87.61%
	Crisis	5.85%	6.73%	87.42%
	Post-crisis	5.69%	5.89%	88.42%
EMO	Whole sample	6.35%	5.88%	87.77%
	Pre-crisis	6.44%	5.95%	87.61%
	Crisis	6.73%	5.85%	87.42%
	Post-crisis	5.89%	5.69%	88.42%

Table 4. The percentage of identically and differently classified trades for the pairs (TR, LR) and (QR, EMO) during the whole sample period from January 2, 2005 to December 30, 2016, for the whole group of 105 WSE-listed companies.

	(TR, LR)	(QR, EMO)
Both unclassified	1.79%	88.84%
The same classification	97.02%	0.04%
Mixed classification (includes opposite classification)	1.19%	11.12%

TR and LR procedures perform better than QR and EMO methods. For this reason, subsequent Table 4. includes more details concerning identically and differently classified trades for the pairs (TR, LR) and (QR, EMO). One can observe that QR and EMO procedures work substantially worse than TR and LR on the WSE.

According to the literature, the main way of assessing accuracy of trade side classification procedures is to compare classification results with true trade directions. The problem with transaction data availability is vast for the WSE, and hence the following question has been formulated:

- Is it possible to directly test the accuracy of trade side classification algorithms for the intraday data from the Warsaw Stock Exchange?

Regardless of the choice of a formal definition of a trade initiator, assigning a trade direction requires presence of the trades and quotes that precede the trade. However, the transaction dataset for the WSE contains only trade prices, the best-bid and best-ask prices, volume, date, and order-entry time[9]. Unlike that, transaction data sets that are utilized in other studies contain more useful information. For example, Aitken and Frino (1996) test the accuracy of the tick rule for the data from the ASX, which is (like the WSE) a pure order-driven market (see Table 2.). They use the data that include all trades and quotes. Lu and Wei (2009) employ database provided by the TWSE, which is also an order-driven market (see Table 2.). The transactions data are preserved in the order, trade, and disclosure files. The trade file contains the date, trade time, order type (buy or sell) of transaction, trade volume, trade serial number, trade price, trade categories, and the identify of a trader. The order file includes the date, stock code, order type (buy or sell), and so on.

Comparing the available data with those used in other studies we conclude that the answer to the above question is rather negative. Unfortunately, we are not able to recognize true trade directions on the WSE. Therefore, we have to rely on the literature and utilize indirect methods for inferring the initiator of a trade in the research concerning various aspects of microstructure of the Polish stock market.

Conclusion

Analysing the accuracy of the trade side classification is obviously important as this accuracy determines the validity of empirical research based on classification methods. In practice, to measure both market liquidity and dimensions of market liquidity based on intraday data, it is essential to recognize the side that initiates a transaction, and to distinguish between the so-called buyer- and seller-initiated trades. However, the problem with transaction data

[9]Transaction data coming from Bloomberg under the license agreement between Bloomberg and Bialystok University of Technology (the grant No. 2016/21/B/HS4/02004).

availability is crucial. Our empirical results based on intraday data rounded to the nearest second indicate that the tick rule and the Lee-Ready algorithm outperforms the quote rule and EMO procedure on the WSE. The main advantage of the tick rule is that it requires only trades (quotes are not necessary). The advantage of the Lee-Ready algorithm is that it combines tick and quote rules, and therefore incorporates more information since it utilizes past quotes. Inasmuch as the TR and LR procedures perform well and the results are qualitatively and quantitatively much the same for the data from the WSE, we can essentially recommend both methods for further research on Polish equity market. Moreover, the empirical experiments confirm that the trade side classification results turn out to be robust to the choice of the period. Specifically, they are not significantly different during the crisis period compared to other investigated periods. The robustness of the TR and LR algorithms to a sample choice is certainly an important merit.

Acknowledgements

This research has been supported by the grant No. 2016/21/B/HS4/02004 from the National Science Centre in Poland.

References

Aitken, M. & Frino, A. (1996). The accuracy of the tick test: Evidence from the Australian Stock Exchange. *Journal of Banking and Finance, 20*, 1715-729.

Asquith, P., Oman, R. & Safaya, C. (2010). Short sales and trade classification algorithms. *Journal of Financial Markets, 13(1)*, 157-173.

Chakrabarty, B., Pascual, R. & Shkilko, A. (2015). Evaluating trade classification algorithms: Bulk volume classification versus the tick rule and the Lee-Ready algorithm. *Journal of Financial Markets, 25*, 52-79.

Easley, D., López-de Prado, M. & O'Hara, M. (2013). Bulk classification of tradingactivity. *Johnson School Research Paper 8-2012*.

Ellis, K., Michaely, R. & O'Hara, M. (2000). The accuracy of trade classification rules: Evidence from Nasdaq. *Journal of Financial and Quantitative Analysis, 35(4)*, 529-551.

Finucane, T. (2000). A direct test of methods for inferring trade direction from intraday data. *Journal of Financial and Quantitative Analysis, 35(4)*, 553-576.

Lee, C.M.C. & Radhakrishna, B. (2000). Inferring investor behavior: Evidence from TORQ data. *Journal of Financial Markets, 3*, 83-111.

Lee, C.M.C., & Ready, M. (1991). Inferring trade direction from intraday data. *Journal of Finance, 46(2),* 733-746.

Lu, Y.C. & Wei, Y.C. (2009). Classification of trade direction for an equity market with price limit and order match: Evidence from the Taiwan stock market. *Investment Management and Financial Innovations, 6(3),*135-147.

Odders-White, E. (2000). On the occurrence and consequences of inaccurate trade classification. *Journal of Financial Markets, 3*, 259-286.

Nowak, S. (2017). Order imbalance indicators in asset pricing: Evidence from the Warsaw Stock Exchange. In: Jajuga, K, Orlowski, L and Staehr, K (eds), *Contemporary Trends and Challenges in Finance. Springer Proceedings in Business and Economics*, Springer, Cham, 91-102.

Olbryś, J. (2017). Interaction between market depth and market tightness on the Warsaw Stock Exchange: A preliminary study. In: Jajuga, K, Orlowski, L andStaehr, K (eds),*Contemporary Trends and Challenges in Finance. Springer Proceedings in Business and Economics*, Springer, Cham, 103-111.

Olbryś, J. & Majewska, E. (2015). Bear market periods during the 2007-2009financial crisis: Direct evidence from the Visegrad countries. *Acta Oeconomica, 65(4)*, 547-565.

Olbryś, J. & Mursztyn, M. (2015). Comparison of selected trade classification algorithms on the Warsaw Stock Exchange. *Advances in Computer Science Research, 12*, 37-52.

Olbryś, J. & Mursztyn, M. (2017). Measurement of stock market liquidity supported by an algorithm inferring the initiator of a trade. *Operations Research and Decisions, 27(4),* 111-127.

Theissen, E. (2001). A test of the accuracy of the Lee/Ready trade classification algorithm. *Journal of International Financial Markets, Institutions and Money, 11(2)*, 147-165.