Technical Document
Market Microstructure Database Xetra

Thomas Johann[1], Stefan Scharnowski[1,2], Erik Theissen*[1],
Christian Westheide[1,2] and Lukas Zimmermann[1]

[1]University of Mannheim
[2]Research Center SAFE (Goethe University Frankfurt)

This version: October 26, 2018

Abstract

In this document, we describe how we construct the Market Microstructure Database Xetra. We obtain trade as well as bid and ask data from Deutsche Börse Group. We apply several filters and compute various market microstructure measures at a daily frequency. We then discuss technical details and potential data problems.

*Corresponding author: University of Mannheim, Finance Area, L9,1-2, 68131 Mannheim, Germany. theissen@uni-mannheim.de

We gratefully acknowledge financial support from the German Science Foundation (DGF) under grant TH 724/6-1.

1 Introduction

Liquidity of financial markets is important. It affects a multitude of economic outcomes such as trading profits, asset prices, real investment decisions, and corporate actions. However, measuring liquidity is no easy endeavor. One problem faced by researchers interested in liquidity is that measuring it oftentimes requires intraday high-frequency data. Such datasets are expensive and, because of their size, hard to work with.

We address this problem for the German equity market by providing a database that contains various daily market-microstructure measures of liquidity on the stock level. Using the Market Microstructure Database Xetra (MMDB-Xetra)[1] allows researchers to focus on their research question without having to measure liquidity themselves.

In this document, we explain how we construct the database and answer questions that might arise while using it. Section 2 describes what stocks are included in the database, section 3 introduces the raw data that we then filter as described in section 4. In section 5, we list and explain all variables contained in the database. Section 6 provides definitions of some measures of liquidity.

While this document is the technical description of the construction of this database, in Johann et al. (2018) we use this data to show recent developments of liquidity in the German stock market. There we provide summary statistics and graphs of the data and also give suggestions on how to further filter the data.

2 Sample

We obtain historical CDAX constituents from Datastream, Compustat, and Bloomberg. Those three lists are not identical. We therefore keep all companies contained in any of those lists. We use Bloomberg's corporate actions calendar to track ISIN changes and dates of companies' IPO and listing/delisting. Finally, we manually collect information on those firms not covered by Bloomberg. The final sample encompasses 982 firms.

[1]"Xetra" and "CDAX" are a registered trademarks of Deutsche Börse AG.

3 Data Source

We obtain time-sorted csv-files containing data for the entire Xetra universe from Deutsche Börse Group.[2] Trades and quotes are in separate files. Trade data is available from January 1999, quote data only from February 2002. The available data fields in the raw data are:

Quote files

1. WKN
2. ISIN
3. INSTRUMENT_NAME
4. TIMESTAMP: format YYYY-MM-DD HH:MM:SS
5. HSEC: subdivides each second into its hundredth
6. PRICE
7. UNITS
8. BID_ASK_FLAG: A=ask, B=bid, I=indicative

Trade files

1. EXCHANGE: EDE = Xetra, EDF = Frankfurt, EDV = EEX (from 8/2000 to 7/2010)
2. DATE: format YYYY-MM-DD
3. TIMESTAMP
4. HSEC
5. WKN
6. ISIN
7. INSTRUMENT_NAME
8. PRICE
9. PRICE_TYPE: price addendum
10. TRADED_UNITS
11. RATE_PRICE_TYPE (from 02/2010): contains information on the source of the trade (e.g. opening auctions, continuous trading,...)

[2]For more details visit Deutsche Börse Data Shop: Xetra Best bid / Best ask data and Xetra and Frankfurt Trading Floor All Traded Instruments Tick data

4 Data Filtering

Each year we combine quote and trade data for each asset identified by its ISIN. We then apply several filters to the intraday data to remove extreme outliers and data errors. We collect the number of observations deleted by the filters per day. Before filtering we have 3.222 bn firm-level intraday observations. The filters in total remove 101 mn observations (3.14% of the sample). Eventually we are left with 3.121 bn firm-level intraday observations.

Within the raw data, some data is missing or obviously erroneous. We set observations with obvious errors to missing. Where possible, we then still compute the intraday measures, even on days when some data is missing. It should be noted that when performing further analyses based on this dataset, those days should be handled with care. We hence flag all such days in the final dataset. Further information on these data problem can be found in Appendix B.

The filters are applied in the following order:

1. Keep only trades from exchange "EDE" (Xetra)[3]
2. Drop indicative quotes and auctions
3. Drop data lacking information on whether it was a bid quote, an ask quote, or a trade.
4. Drop observations outside the trading hours[4]
5. Drop trades if trade price or volume are not strictly positive
6. Drop trades if the trade price is six standard deviations of daily trade prices below (above) the minimum (maximum) of the opening price, mean daily price or closing price.
7. Eliminate price spikes: If the transaction price of a trade is 20 percent plus 1 Euro above the price of the previous trade and the following trade, the trade is excluded. If the transaction price of a trade is 20 percent minus 1 Euro below the price of the previous and the following trade, the trade is excluded.
8. Drop quotes if the ask or bid price is zero or negative.
9. Drop quotes if the ask price is less or equal to the bid price.
10. Drop quotes if the depth is not strictly positive.

[3]The other exchange in the data is the Frankfurt trading floor (Parkett).
[4]See appendix A for the exact trading hours.

11. Drop quotes if the quote midpoint is 50 Euro or less and the quoted half-spread is greater than 2.50 Euro; drop quotes if the quote midpoint is greater than 50 Euro and the quoted half-spread is greater than 5 percent.
12. Drop quotes if the bid price of a quote is less than the minimum of the first bid, the average bid, or the last bid price of the stock-day minus six standard deviations of the bid price. Drop quotes if the ask price of a quote is more than the maximum of the first ask, the average ask, or the last ask price of the stock-day plus six standard deviations of the ask price.

5 Variable definitions

We construct the following variables for each day. To keep the notation concise, we sometimes use ***[placeholders]*** that can take multiple values.

Identifiers and time variables

1. **wkn**: WKN
2. **instrument_name**: Instrument name
3. **final_isin**: ISIN as of the end date of the sample period
4. **isin**: ISIN as of the respective date
5. **year**: Year
6. **month**: Month
7. **day**: Day
8. **data_problem**: Unsolvable data issues with the raw data [5]

Sample size before filtering

9. ***[trades/bids/asks]*****_start**: Total number of trades/bids/asks before any filter is applied
10. **indicative_start**: Total number of indicative quotes before any filter is applied
11. **trades_from_other_exchanges**: Total number of trades from other exchanges than Xetra (before filter 1)

[5] Appendix B describes the data problems and the flags used in this variable in detail.

Auctions

We differentiate between the four kinds of auctions. In what follows, $\star$ is a placeholder for either of the four auction types.

- An *open* auction is an opening auction
- An *intra* auction is a scheduled intraday auction
- A *close* auction is a closing auction
- A *vola* auction is any auction after a volatility interruption. There might be multiple such auction on a given day.

12. **auctionprice_$\star$**: The price of a $\star$-auction. In case of multiple *vola* auctions, their mean price is given.

13. **auctionunit_$\star$**: The volume, measured as the number of shares, of a $\star$-auction. In case of multiple *intra* or *vola* auctions, their total volume is given.

14. **total_auctions**: Total number of auctions

15. **indicative_quotes**: Total number of indicative quotes (before filter 2)

Number of observations dropped due to filters 3–12

16. **no_indicator**: Total number of lines neither flagged as trade or quote (before filter 3)

17. ***[trade/bid/ask]*_before_trading_hours**: Total number of trades/bids/asks before trading hours (before filter 4)

18. ***[trade/bid/ask]*_after_trading_hours**: Total number of trades/bids/asks after trading hours (before filter 4)

19. **negative_or_zero_trade_price**: Total number of negative or zero price trades (before filter 5)

20. **negative_or_zero_trade_volume**: Total number of negative or zero volume trades (before filter 5)

21. **below_min**: Number of trades priced at more than 6 standard deviations below the min of the mean/opening/closing price (before filter 6)

22. **above_max**: Number of trades priced at more than 6 standard deviations above the max of the mean/opening/closing price (before filter 6)

23. **up_spike**: Number of trades priced at more than 20% + 1 Euro above the previous and following trades' price (before filter 7)

24. **down_spike**: Number of trades priced at more than 20% + 1 Euro below the previous and following trades' price (before filter 7)

25. **negative_or_zero_*[bid/ask]*_price**: Total number of negative or zero bid/ask price quotes (before filter 8)

26. **zero_spread**: Total number of zero spread quotes (before filter 9)

27. **negative_spread**: Total number of negative spread quotes (before filter 9)

28. **negative_or_zero_depth_*[bid/ask]***: Total number of negative or zero bid/ask volume quotes (before filter 10)

29. **large_spread**: Total number of spread observations eliminated by filter 11

30. **below_min_*[bid/ask]***: Number of bid/ask quotes priced at more than 6 standard deviations of trade prices below the min mean/opening/closing bid/ask quote of that day (before filter 12)

31. **above_max_*[bid/ask]***: Number of bid/ask quotes priced at more than 6 standard deviations of trade prices above the max mean/opening/closing bid/ask quote of that day (before filter 12)

The following variables (32–79) are calculated based on the sample remaining after application of all filters.

Opening and closing quotes

32. **opening_*[bid/ask]*_price**: Price of the first bid/ask

33. **opening_*[bid/ask]*_volume**: Volume of the first bid/ask

34. **opening_*[bid/ask]*_value**: Value (price*volume) of the first bid/ask

35. **closing_*[bid/ask]*_price**: Price of the last bid/ask

36. **closing_*[bid/ask]*_volume**: Volume of the last bid/ask

37. **closing_*[bid/ask]*_value**: Value (price*volume) of the last bid/ask

Opening and closing trades

38. **opening_trade_price**: Price of the first trade (non-auction trades)

39. **opening_trade_volume**: Volume of the first trade(non-auction trades)

40. **opening_trade_value**: Value (price*volume) of the first trade (non-auction trades)
41. **closing_trade_price**: Price of the last trade (non-auction trades)
42. **closing_trade_volume**: Volume of the last trade (non-auction trades)
43. **closing_trade_value**: Value (price*volume) of the last trade(non-auction trades)

Minimum and maximum trade prices

44. **max_price**: Maximum trade price
45. **min_price**: Minimum trade price

Average quotes, updates and improvements

46. ***[bid/ask]*_time**: Time-weighted bid/ask price
47. ***[bid/ask]*_updates**: Number of bid/ask updates
48. ***[bid/ask]*_improvements**: Number of bid/ask improvements

Quoted spreads

49. **quoted_time**: Time-weighted *Quoted Spread*[6]
50. **rel_quoted_time**: Time-weighted *Relative Quoted Spread*

Quoted depth

51. ***[bid/ask]*_depth_time**: Time-weighted depth on the bid/ask side.
52. ***[bid/ask]*_depth_euro_time**: Time-weighted depth on the bid/ask side in terms of value (units*price)

Volatility measures

53. **midpoint_changes**: Total number of *Midpoint* changes, where midpoint is the last mipoint in a 5-minute interval
54. **midpoint_vola**: *Midpoint Volatility*

Number, volume, and value of trades (continuous trading)

[6] Italic terms are defined in section 6

55. **final_buys**: Total number of buyer-initiated trades[7]

56. **final_sells**: Total number of seller-initiated trades

57. **final_trades**: Total number of trades

58. **sum_sell_val**: Total selling value (price*volume)

59. **sum_sell_vol**: Total selling volume

60. **sum_buy_val**: Total buying value (price*volume)

61. **sum_buy_vol**: Total buying volume

Order imbalance

62. **oib**: *Order Imbalance*

63. **oib_val**: *Order Imbalance* in terms of value (price*volume)

64. **oib_vol**: *Order Imbalance* in terms of volume

65. **lambda**: *Adverse Selection Component Lambda* as in Lin et al. (1995)

The following variables (66–79) are aggregated as volume-weighted and equally-weighted means over all trades during continuous trading. The variable names contain the suffixes ***value*** and ***equal***, respectively.

Trade prices and volume

66. **trade_price_*[value/equal]***: Trade price

67. **trade_price_*[buy/sell]*_*[value/equal]***: Trade price if trade is a buy/sell

68. **trade_qty_equal**: Average trade volume

Trading volume and value

69. **buy_vol_equal**: Average volume traded in buyer-initiated trades

70. **sell_vol_equal**: Average volume traded in seller-initiated trades

71. **buy_val_equal**: Average value (price*volume) traded in buyer-initiated trades

72. **sell_val_equal**: Average value (price*volume) traded in seller-initiated trades

[7] We use Lee and Ready (1991)'s algorithm to classify trades into buys and sells.

Quote midpoints

73. **midpoint_*[value/equal]***: *Midpoint*

Effective spreads and price impact

74. **effective__*[value/equal]***: *Effective Spread*
75. **rel_effective__*[value/equal]***: *Relative Effective Spread*
76. **pi_*__*[value/equal]***: *Price Impact* (* is 5 or 60 [minutes])
77. **pi_kyle_*__*[value/equal]***: *Price Impact as in Kyle* (* is 5 or 60 [minutes])

Depth

78. ***[bid/ask]*_depth__*[value/equal]***: Depth on the bid/ask side
79. ***[bid/ask]*_depth_euro__*[value/equal]***: Depth in terms of value (price*units) on the bid/ask side

6 Construction of Liquidity Measures

We use the index i for a firm $i \in I$ (total number of firms), $t \in T$ for a trading day, the index $j \in J$ identifies a trade or the corresponding quote immediately before the trade, where J is the total number of trades, and $\psi \in \Psi$ as the number of the quote update, where Ψ is the total number of quote updates for a stock i on day t.

- **Quoted Spread**:

$$quoted_{it\psi} = ask_{it\psi} - bid_{it\psi} \tag{1}$$

- **Midpoint**:

$$M_{it\psi} = \frac{bid_{it\psi} + ask_{it\psi}}{2}. \tag{2}$$

- **Relative Quoted Spread**:

$$rel\_quoted_{it\psi} = \frac{ask_{it\psi} - bid_{it\psi}}{M_{it\psi}} = \frac{quoted_{it\psi}}{M_{it\psi}} \tag{3}$$

- **Effective Spread**:

$$effective_{itj} = 2 * |P_{itj} - M_{itj}| \tag{4}$$

- **Relative Effective Spread**:

$$rel\_effective_{itj} = \frac{2 * |P_{itj} - M_{itj}|}{M_{itj}} = \frac{effective_{itj}}{M_{itj}} \tag{5}$$

- **Midpoint Volatility**:

  Let M_{itf}^{5min} be the last midpoint of a five minute interval f and F the total number of daily five minute intervals without missing midpoints.

$$midpoint\_vola_{it} = \frac{1}{F-1} \sum_{f=2}^{F} \left(\frac{M_{itf}^{5min} - M_{it(f-1)}^{5min}}{M_{it(f-1)}^{5min}} \right)^2 \tag{6}$$

- **Price Impact**:

  Let M_{itj}^{+5min} be the midpoint five minutes after M_{itj}, the midpoint immediately before a transaction at time t, and let

$$q_{itj} = \begin{cases} 1 & \text{if trade } itj \text{ is a buy} \\ -1 & \text{if trade } itj \text{ is a sell} \end{cases}$$

Then the price impact is defined as

$$price\_impact_{itj} = \frac{\left(M_{itj}^{+5min} - M_{itj}\right) * q_{itj}}{M_{itj}} \tag{7}$$

$$price\_impact_{itj}^{kyle} = \frac{\left(M_{itj}^{+5min} - M_{itj}\right) * q_{itj}}{M_{itj} * volume_{itj}} \tag{8}$$

Where $volume_{itj}$ is the number of shares traded in a transaction. We also calculate these variables for a lag of 1 and 60 minutes.

- **Order Imbalance**:

  Order imbalance is the difference between buyer and seller initiated transactions relative to all transactions, computed using either the number of transactions, the number of shares traded in these transactions, or the value of these transactions:

$$OIB_{it} = \frac{final\_buys_{it} - final\_sells_{it}}{final\_buys_{it} + final\_sells_{it}} \tag{9}$$

$$OIB_{it}^{vol} = \frac{sum\_buy\_vol_{it} - sum\_sell\_vol_{it}}{sum\_buy\_vol_{it} + sum\_sell\_vol_{it}} \tag{10}$$

$$OIB_{it}^{val} = \frac{sum\_buy\_val_{it} - sum\_sell\_val_{it}}{sum\_buy\_val_{it} + sum\_sell\_val_{it}} \tag{11}$$

- **Lambda**:

  We estimate the adverse selection component λ_{it} of the effective spread using the methodology of Lin et al. (1995). λ_{it} is obtained from the following regression, which is repeated for each firm and day.

$$\log\left(\frac{M_{itj}}{M_{it(j-1)}}\right) = \lambda_{it} * \log\left(\frac{P_{it(j-1)}}{M_{it(j-1)}}\right) + \epsilon_{itj} \tag{12}$$

- **Depth Ask/Bid Side**: *Depth* on the ask/bid side is defined as the number of shares available at the best bid or ask.

$$\begin{aligned} d_{it\psi}^{ask} &= ask\_quantity_{it\psi} \\ d_{it\psi}^{bid} &= bid\_quantity_{it\psi} \end{aligned} \tag{13}$$

- **Depth Ask/Bid Side in Terms of Value**: *depth_euro* is then calculated as

$$\begin{aligned} d_{it\psi}^{val,ask} &= ask\_quantity_{it\psi} * ask\_price_{it\psi} \\ d_{it\psi}^{val,bid} &= bid\_quantity_{it\psi} * bid\_price_{it\psi} \end{aligned} \tag{14}$$

References

Johann, T., S. Scharnowski, E. Theissen, C. Westheide, and L. Zimmermann (2018). Liquidity in the German stock market. *Working paper*.

Lee, C. M. C. and M. J. Ready (1991). Inferring trade direction from intraday data. *The Journal of Finance 46*(2), pp. 733–746.

Lin, J.-C., G. C. Sanger, and G. G. Booth (1995). Trade size and components of the bid-ask spread. *The Review of Financial Studies 8*(4), pp. 1153–1183.

Appendix

A Trading Hours

- from 28/11/1997: 08.30h – 17.00h
- from 20/09/1999: 09.00h – 17.30h
- from 02/06/2000: 09.00h – 20.00h
- from 03/11/2003: 09.00h – 17.30h
- last trading day of the year: until 14.00h

B Data Problems

Table B1 describes data problems with the original data. The column *Problem* represents the issue stored in the variable data_problem and corresponds to the flags used.

Table B1: Data Problems

Problem	From	Until	Obs	% of Sample
Ask amount equals ask price	05/09/2002	19/11/2002	33,081	1.65
No quotes	01/09/2011	22/09/2011	6,499	0.33
No trade data	19/06/2012	19/06/2012	433	0.02
No quotes (for parts of the sample)	01/09/2012	30/09/2012	2,804	0.14
Missing indicative quotes	26/11/2012	30/04/2013	15,248	0.76
All			59,189	2.96

C Details on classifying auctions

It is important to separate trading activity during auctions and during continuous trading because auction-related trades tend to be much larger than other trades. If not appropriately dealt with, including these auction-related trades when calculating intraday measures of liquidity might bias the results. The raw data, however, only clearly identifies auction trades starting February 2012.

There are four kinds of auctions on Xetra:

- **Opening auctions** occur at around 9:00h.
- Scheduled **intraday auctions** occur at around 13:00h. Between 2002 and November 2003 they can additionally occur at around 17:30h. On the last trading day of each year, there is no scheduled intraday auction.
- **Closing auctions** generally occur at around 17:30h, or between 2002 and November 2003 at around 20:00h. On the last trading day of the year, the closing auction takes place at around 14:00h.
- **Volatility auctions** to restart trading after volatility interruptions. There can be more than one volatility interruption per stock and day.

We hence need a way to identify all auction-related quotes (indicative quotes[8]) and trades for these types of auctions. While indicative quotes are clearly marked in the raw data, flags that indicate if a trade belongs to an auction (and what kind of auction) are only available starting February 2012. Thus, we build an algorithm to determine auction trades. We use the data after February 2012 to benchmark our algorithm. Depending on the liquidity of the stock, we are able to correctly classify 95–99% of all auction trades. As we do not have any quote data before February 2002, we cannot identify any auctions before that date.

The difficulty lies in the fact that auction trades are not necessarily recorded in correct order. Sometimes we observe a block of indicative quotes (suggesting that there currently is an auction), followed by a trade from the continuous trading session and only after that followed by the auction trade. A simple heuristic taking the first trade after an auction thus yields insufficiently accurate results. Likewise, auction trades do not have to be the largest trades of the day (though they often are), so simply taking the largest trade after we observe indicative quotes does not work either. Additionally, not all auctions occur on every day and not every auction results in a positive volume. The number of volatility interruptions and thus auctions

[8]Indicative quotes show the current market clearing price and corresponding volume if the auction were to end at the moment of the quote.

itself is not known and has to be inferred from the data as well.

The algorithm to identify and classify auction-related trades works as follows:

1. First we drop all indicative quotes with zero volume.
2. Then we identify blocks of indicative quotes, allowing for short gaps of up to 15 minutes in between. We consider each block to be a new auction.
3. We define a time window after the block of indicative quotes wherein we search for the auction trade. The maximum window length is one hour for most stocks and four (eight) hours for stocks with up to five (two) trades in total. In some cases the indicative quotes of the opening or intraday auction are recorded at midnight, so we extend the time windows to include times when the corresponding trades are typically recored, i.e., 8:50h – 9:05h for opening auctions and 12:55h – 13:35h for intraday auctions. In any case, the time window ends when the next block of indicative quotes starts.
4. Indicative quotes are sometimes reported out of order, so we do not precisely know which is the final indicative quote that triggers a trade. Hence, we retain the two last indicative quotes and the indicative quote with the largest volume.
5. For each block of indicative quotes, we then search for the trade within the specified time window that most closely matches either of the three indicative quotes in terms of price and volume. To do so, we minimize the following matching function:

$$\text{Auction Trade}_i = \underset{k\,\in\,Q_i}{\text{argmin}}\left[\underset{j\,\in\,T_{wi}}{\text{argmin}}\left[\left(\frac{P_k - P_j}{P_k + P_j}\right)^2 + \left(\frac{V_k - V_j}{V_k + V_j}\right)^2\right]\right] \tag{15}$$

   where Q_i are the three mentioned indicative quotes of indicative block i, T_{wi} are all trades within time window w_i after the indicative block, and P and V are price and volume of quotes and trades.
6. The trade that minimizes the matching function is then identified as an auction trade, but only if the resulting value of the function is less than 0.4 to avoid matching trades that substantially differ from the indicative quotes.
7. Finally, each identified auction trade is classified as belonging to a specific auction in the following way.
   An **opening auction trade** is an auction trades if it is reported between 8:55h and 9:15h and if the auction is the first on that day.
   A scheduled **intraday auction trade** is an auction trade if it is reported between 12:55h and 13:20h and if it is the first auction trade in that window. Before November

2003, it can also be an auction trade reported between 17:25h and 17:50h if the auction is not the last of the day.

A **closing auction trade** is the last auction trade of the day if reported between 17:25h and 17:50h or after 19:55h between February 2002 and November 2003. On the last trading day of the year, it is the last auction trade if reported between 13:55h and 14:20h.

All remaining auction trades are classified as **volatility auction trades**.