Five Ways Your Historical Data Lie

Common Issues and Misunderstandings with Historical Market Data

July 2021

TABLE OF CONTENTS

Corrected Data Lies

Corrected vs "As-Is" Data

Restrained by costs and lacking technological ability, small market data vendors often choose to obtain data from exchange archives, which have been corrected and cleaned

Corrected Data vs "As-Is" Data

- **What are the differences?**
  - Real-time data is imperfect, subject to exchange publishing mistakes (such as out-of-sequence packets) and trade cancellations;
  - Exchange archived data has been corrected and cleaned, thus is not "as-is".
- **Algorithm Fragility**: Using corrected data for backtesting risks algorithm lacking robustness and breaking in the real-time scenario.

Minute Bar Nuances

- **Trade Cancellation**: Some data vendors deduct cancelled volume from the minute cancellation message received.
- **Condition Codes and their combinations**: Some data vendors use inconsistent inclusion/exclusion rules on trade/quote condition codes.

Open/Close Lies

Consolidate O/C vs Primary Exchange O/C

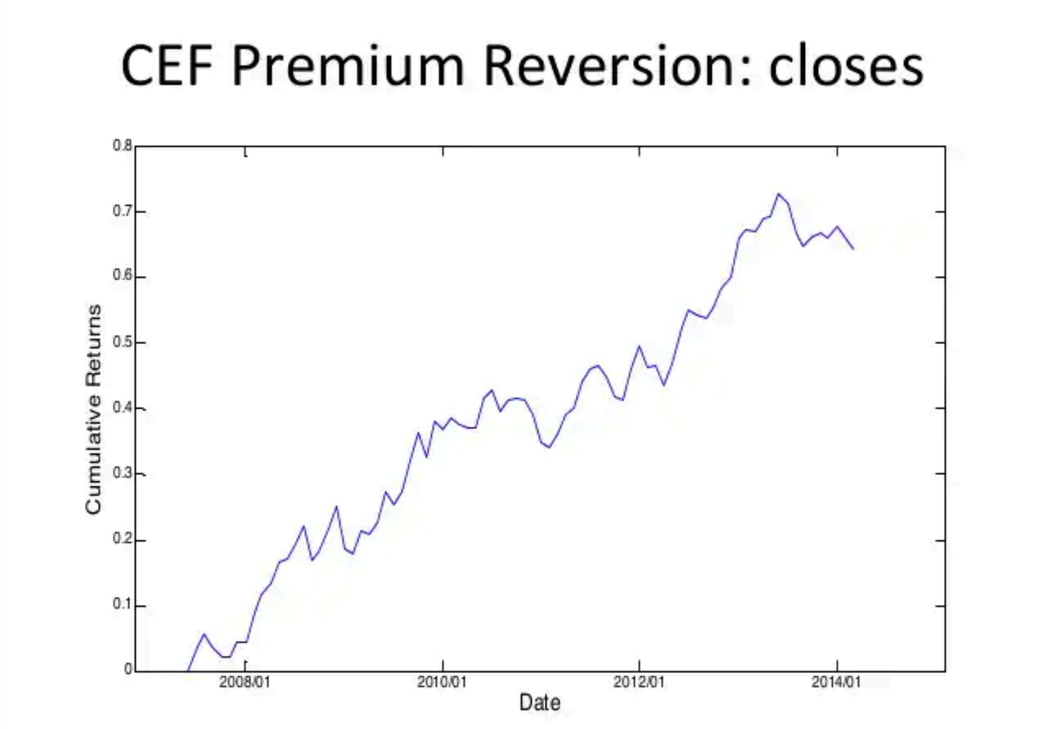
Example 1: CEF Premium Reversion

- Patro *et al* published a paper on trading the mean reversion of closed-end funds' (CEF) premium.
  - ssrn.com/abstract=2468061
- CEFs with high premium (market cap-NAV) will have negative returns, while those with steep discount will have positive returns.
- Rank CEFs based on % premium and buy the bottom quintile and short the top quintile with *monthly* rebalancing.

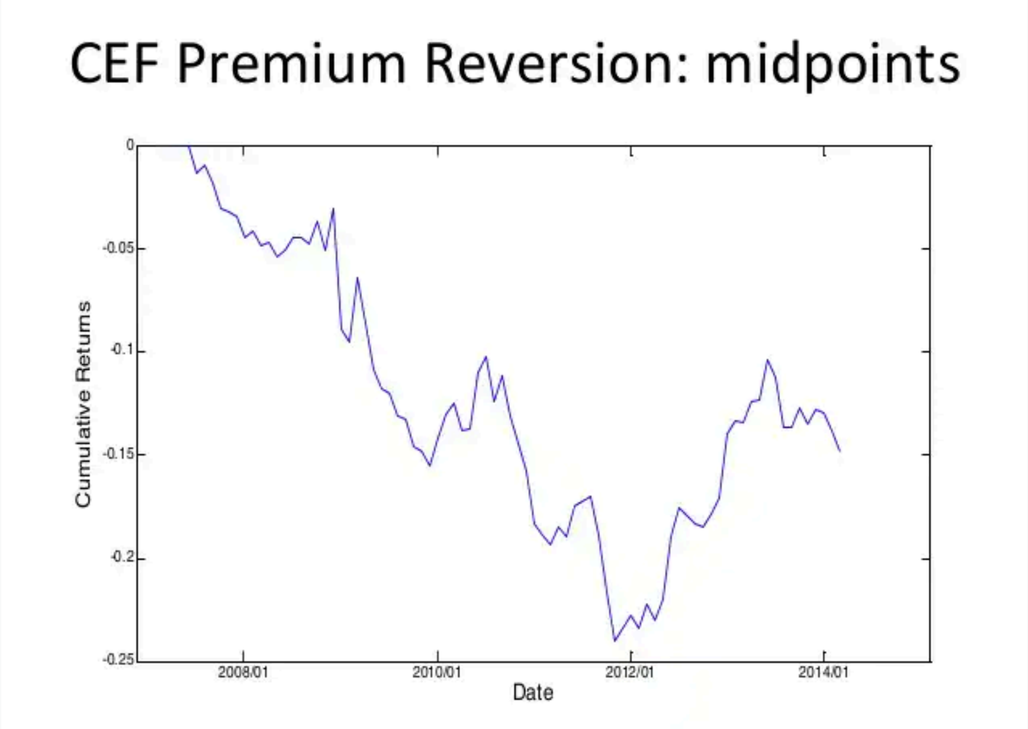
- Author obtained fund price and shares outstanding data from CRSP, and fund NAV from Bloomberg.
- Sharp Ratio is 1.5 from 1998 to 2011
- Dr. Ernie Chan repeated the backtest using CRSP prices and fund NAV from Computstat from 2007 to 2014.

Open/Close Lies

Consolidate vs Primary Exchange O/C



Using CRSP provided closing prices



Using bid/ask midpoints as closing prices

Open/Close Lies

Consolidate vs Primary Exchange O/C

- You wouldn't think using CRSP closing prices vs midpoint between bid/ask at the close matter for a strategy that rebalance only monthly!
- Actual execution will use MOC (Market-on-Close) or LOC (Limit-on-Close) order.
- Actual execution prices will the primary exchange closing auction/cross prices, not the so called "consolidated closing prices" most of market data vendors provide.

algoseek provides Primary Exchange OHLC dataset to help traders avoid

Open/Close Lies

Consolidate vs Primary Exchange O/C

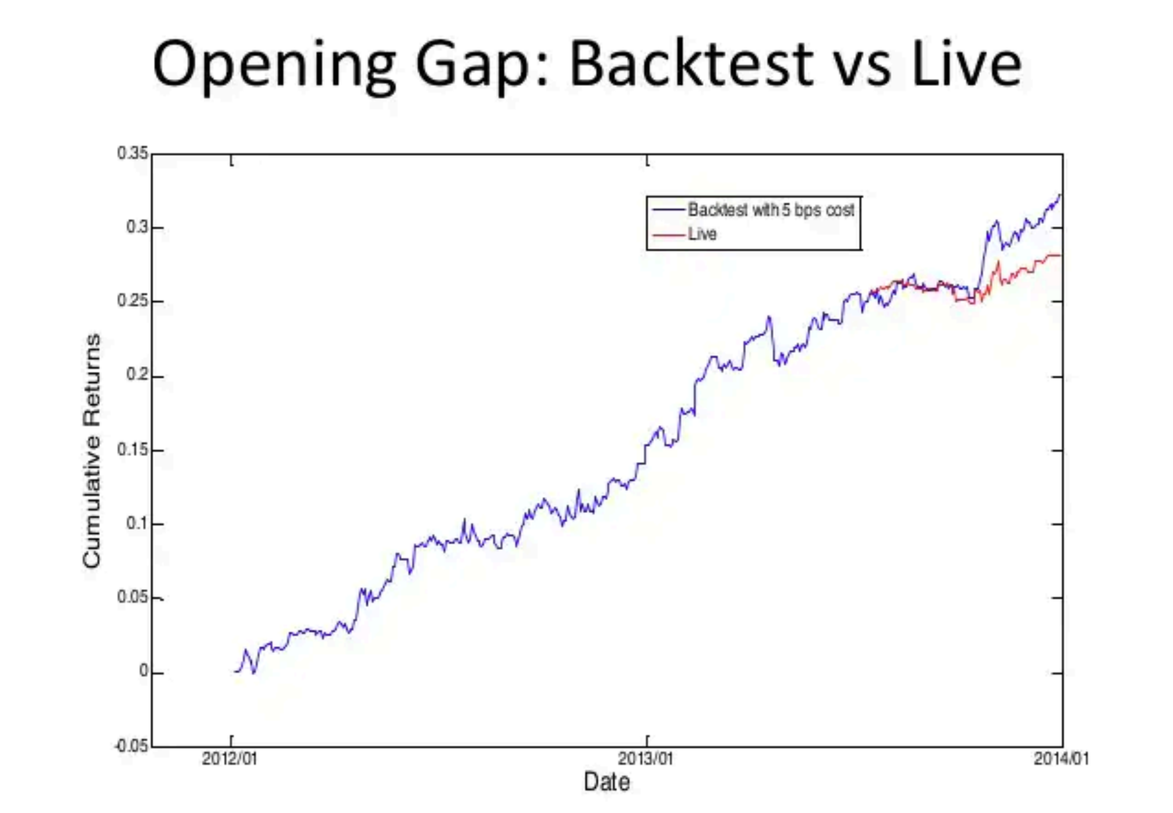
Example 2: Opening gap

- Rank stocks based on their returns from previous close to today's open: retGap.
- Apply fundamental and technical filters e.g. eliminating stocks which just had earnings announcements.
- Buy 10 stocks with the lowest retGap, and short 10 with the highest retGap at the open.
- Exit at the same day's close.
- Backtest from 2012-2014.
- Live trading from mid 2013-2014.

This is another O/C prices related experiment Dr. Ernie Chan conducted in 2015.

Open/Close Lies

Consolidate vs Primary Exchange O/C



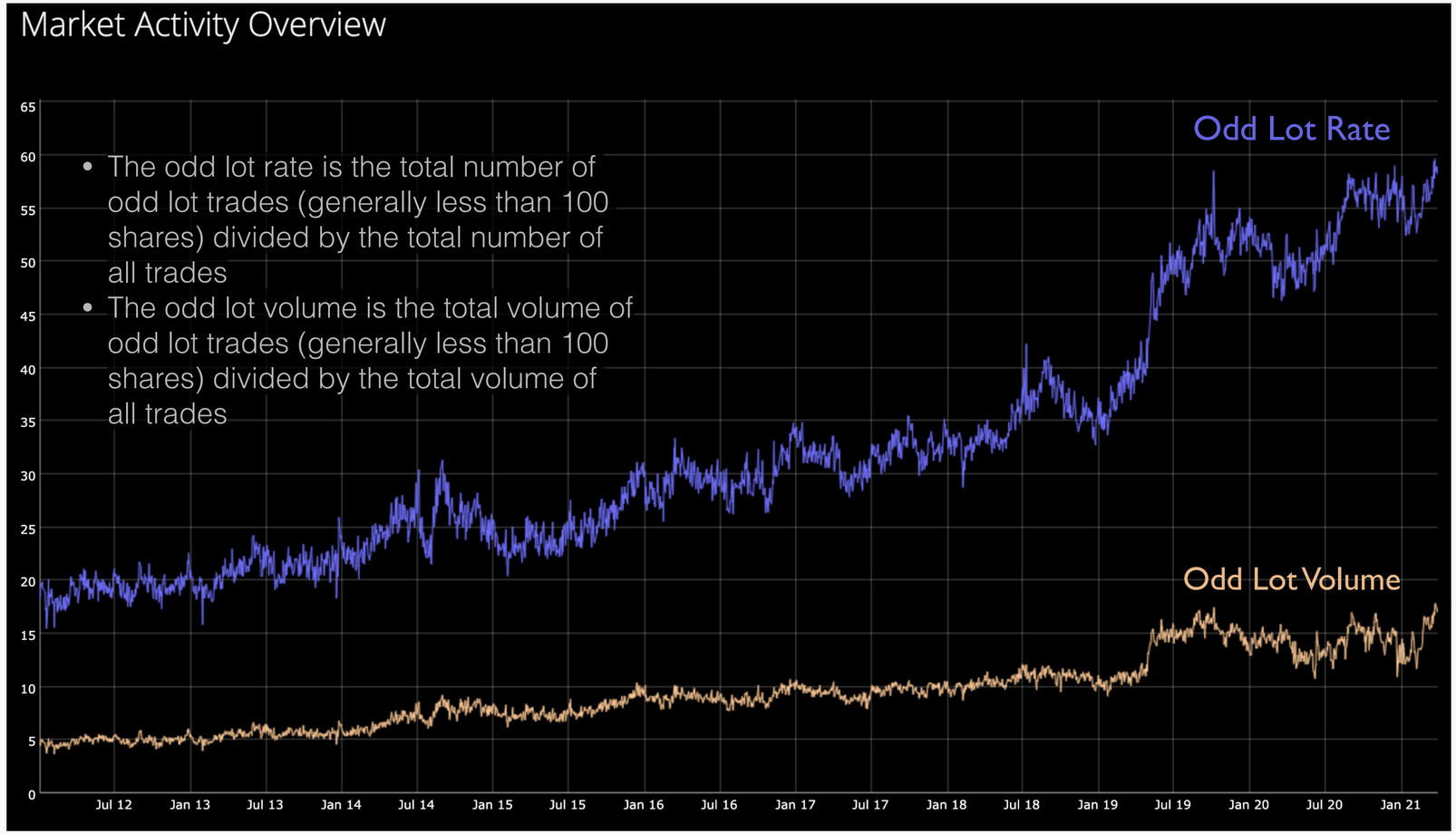
- During backtest, midpoints at close were used (In 2015, it was the closest thing to primary exchange closing price, before algoseek created Primary Exchange OHLC dataset).
- Backtest used CRSP provided opening prices, and also included 5 bps per trade transaction cost.
- Live trading underperformed backtest substantially.
- **Conclusion: open prices also need to use primary exchange auction/ cross prices.**

Why is it so hard for data vendors to provide primary exchange O/C?

- Before 2019, most exchanges did not publish flag showing opening/closing cross trade.
- Historically some exchanges published the Opening Cross as multiple trades instead of just one trade
- All exchanges started using "Official Open" and "Official Close" flags **only since Sep 2019**, and opening prices from non-primary exchanges usually arrive earlier than the official opening price from the primary exchange

algoseek provides Primary Exchange OHLC dataset to help traders avoid the open/close trap.

Tricky Odd Lots

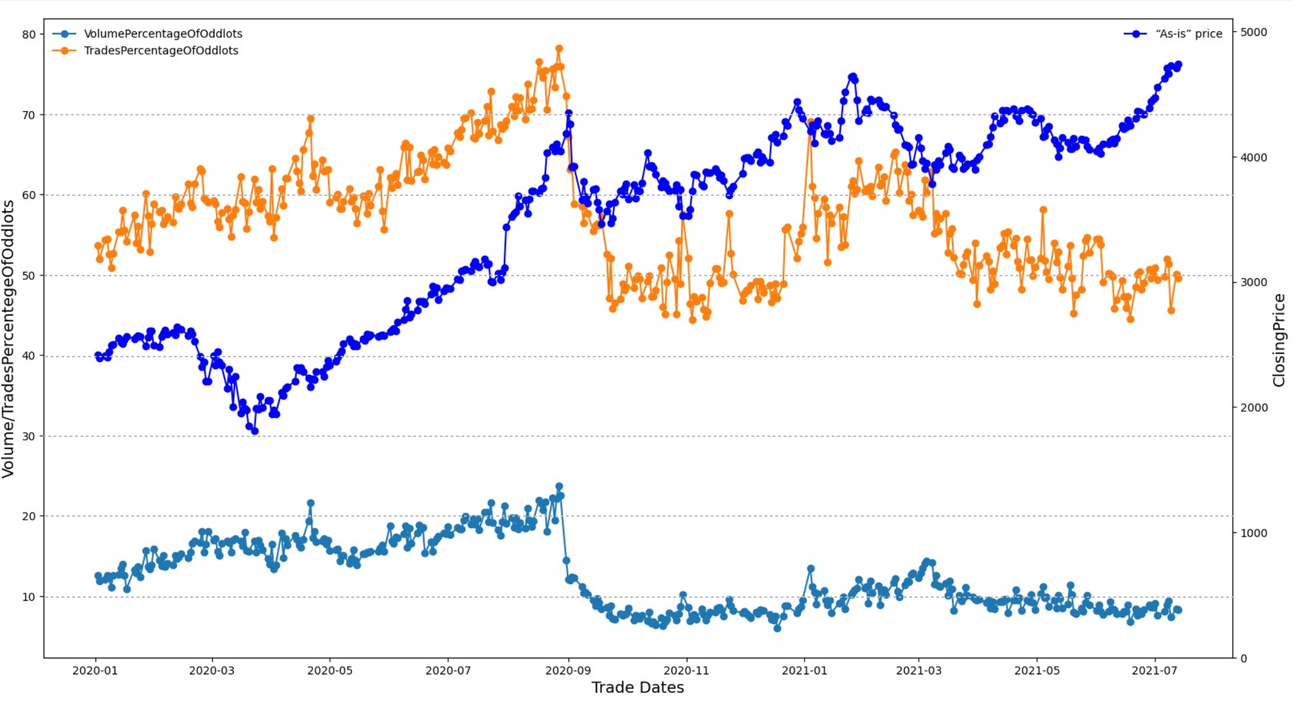


Data Source: SEC Website

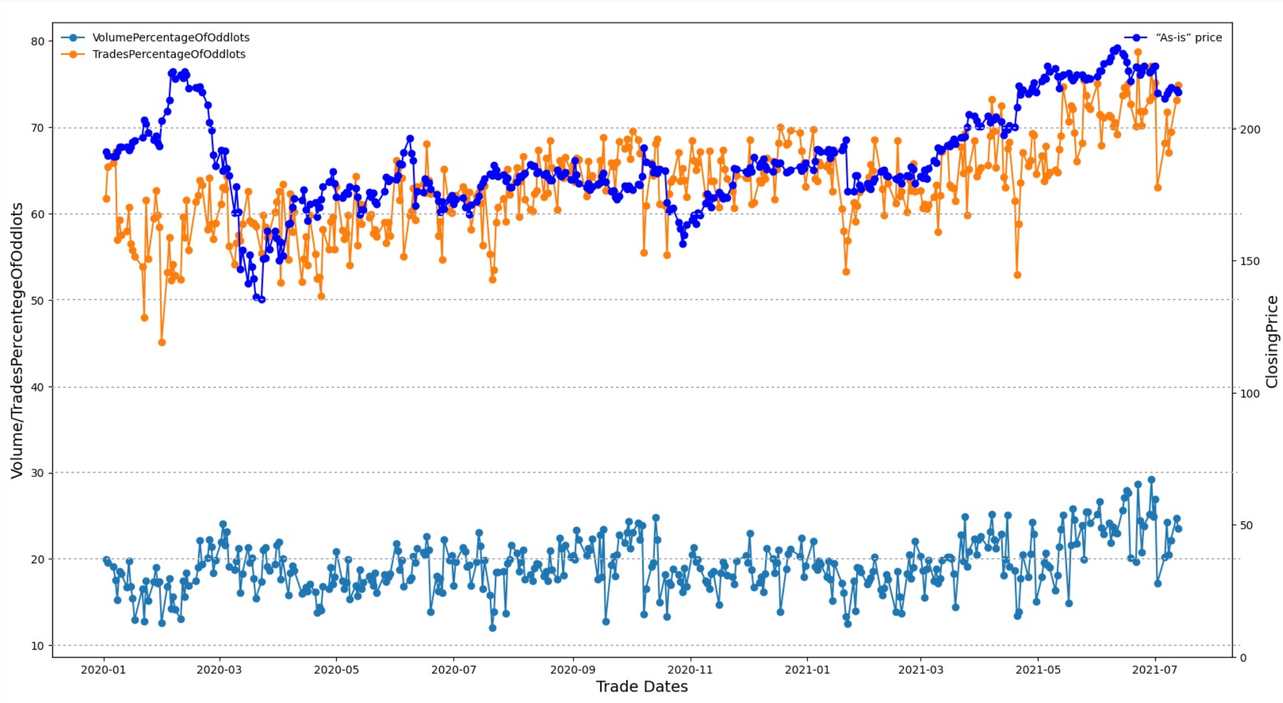
- Odd lot rate and volume have been rising consistently during the past decade.
- In 2021, odd lots account for over 50% of all trades and over 15% of all trade volume/

Odd Lots Activities:
Blue Chip Stocks vs Meme Stocks

AAPL



IBM

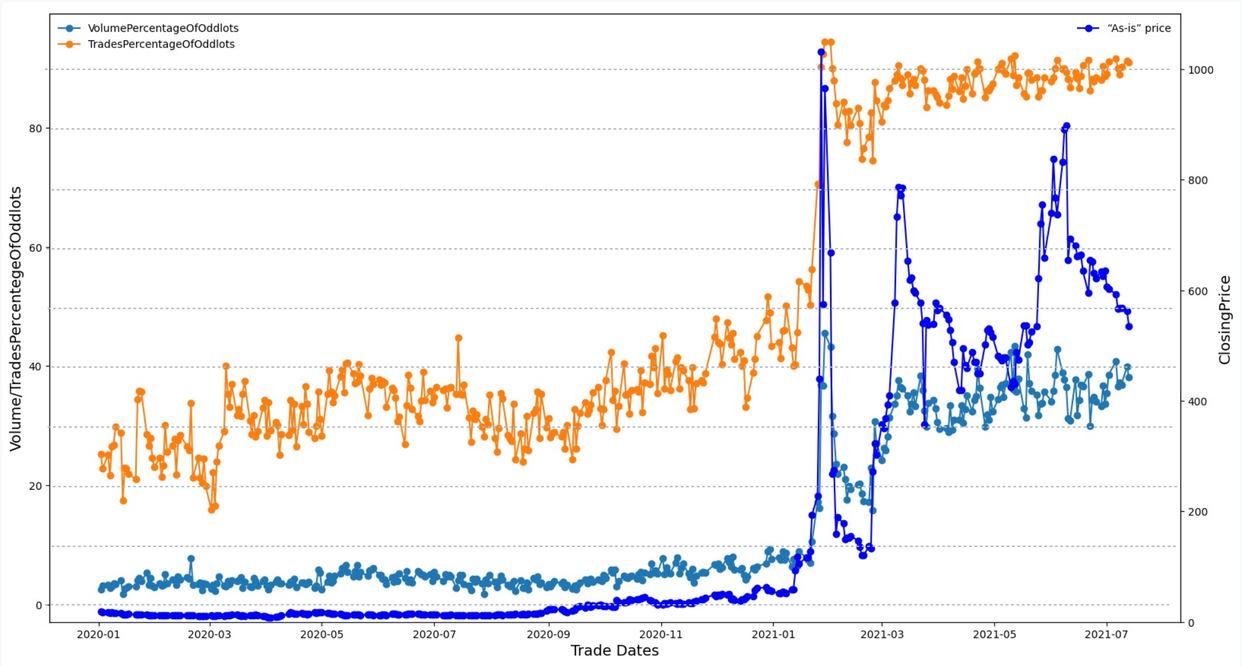


Data Source: algoseek Market Analytics

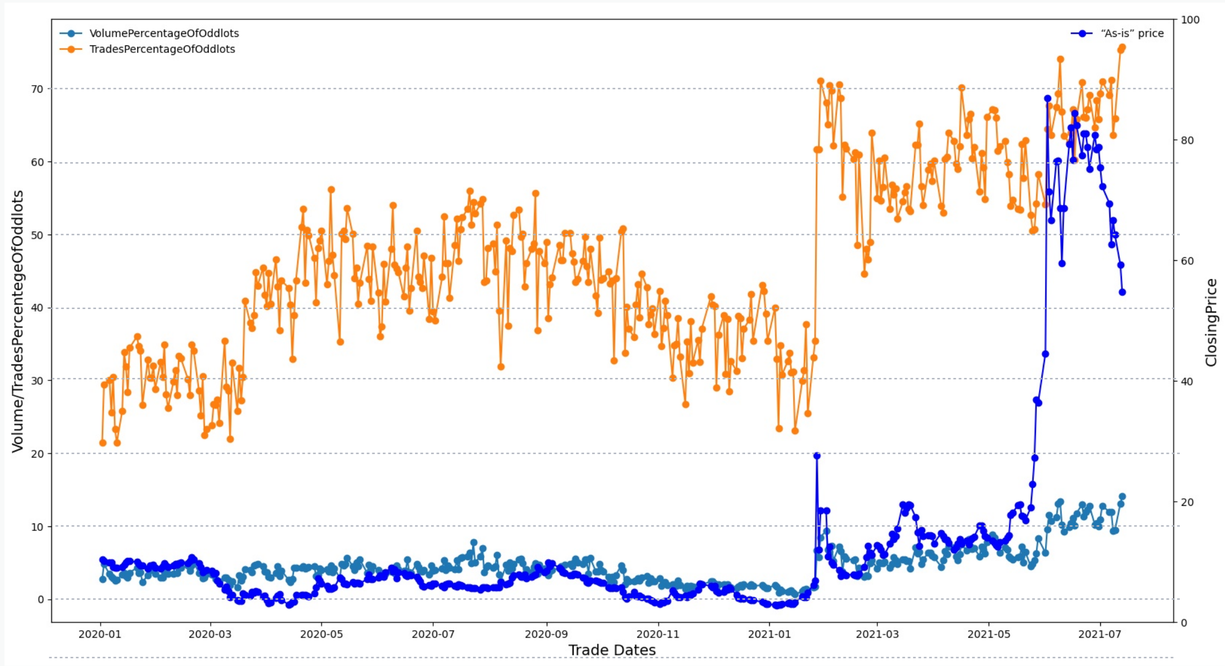
algoseek
THE MARKET DATA COMPANY

Odd Lots Activities:
Blue Chip Stocks vs Meme Stocks

GME



AMC



Data Source: algoseek Market Analytics

algoseek
THE MARKET DATA COMPANY

Odd Lots Lie: False Entries/Exits

- Odd lot orders and TRF trades are not protected by Rule 611 of Regulation NMS, which means they are often executed at prices inferior to NBBO.
- When using minute bar for backtesting price-based entry/exit strategies, irregular trade prices from odd lots often create false entries/exits.

Different Versions of Minute Bars for Different User Needs

- TAQ/Trade Only Minute Bar Version 1:
  - Created from round lots;
  - odd lots & trades from TRF reports are excluded
- TAQ/Trade Only Minute Bar Version 2:
  - Created from only odd lots;
  - For users who want to focus on retail activity
- TAQ/Trade Only Minute Bar Version 3:
  - Include round lots, odd lots and TRF.
  - algoseek TAQ minute bar has over **80 data points** and provides in-depth analytics on market dynamic and microstructure

Your data vendor may lie to you: data quality issues

An algoseek client conducted a comparison between algoseek and vendor R trade-only minute bar data, and yielded astonishing results.

- **7.6% rows were missing** in vendor R's trade-only minute bar
- **High proportion (78.19%) of rows with volume differences** between algoseek and vendor R's minute bars during extended hours

All Trading Hours		Extended Hours		Market Hours	
Field	Percentage(%)	Field	Percentage(%)	Field	Percentage(%)
OpenDiff	0.7	OpenDiff	0.93	OpenDiff	0.33
CloseDiff	0.95	CloseDiff	1.33	CloseDiff	0.36
HighDiff	0.57	HighDiff	0.75	HighDiff	0.29
LowDiff	0.59	LowDiff	0.77	LowDiff	0.32
VolumeDiff	48.63	VolumeDiff	78.19	VolumeDiff	1.0

Example 1:
Missing trades & Wrong Timestamp from Vendor R

Ticker	Date	Bar Start Time	Vendor R Volume	algoseek Volume	Difference	Bloomberg Volume
NXPI	20210125	16:40	50	50	0	50
NXPI	20210125	16:41	2	2	0	2
NXPI	20210125	16:43	6	1778	-1772	1778
NXPI	20210125	16:51	**No Record**	1914	**NA**	1914
NXPI	20210125	16:52	1822	50	+1772	50

Table: Volume Comparison between Vendor R & algoseek Minute Bar (NXPI, Jan 25, 2021)

Note: algoseek volume data is the same with Bloomberg volume data.

Example 1: Missing trades & Wrong Timestamp from Vendor R

Ticker	Date	Time	Event Type	Price	Volume	Exchange	Sales Condition	Note
NXPI	20210125	16:40:26	TRADE	174.37	50	ARCA	Regular, IntermarketSweep, FormT, OddLot, TradeThroughExempt	
NXPI	20210125	16:41:40	TRADE	174.42	1	FINRA	Regular, FormT, OddLot	
NXPI	20210125	16:41:40	TRADE	174.42	1	FINRA	Regular, FormT, OddLot	
NXPI	20210125	16:43:34	TRADE	174.37	6	ARCA	Regular, IntermarketSweep, FormT, OddLot, TradeThroughExempt	
NXPI	20210125	16:43:44	TRADE	174.42	1772	FINRA	Regular, FormT, TradeThroughExempt	**Delayed by Vendor R to 9 minutes later**
NXPI	20210125	16:51:21	TRADE	174.42	1914	FINRA	FormT, TradeThroughExempt, NextDay	**"Lost" by Vendor R**
NXPI	20210125	16:52:07	TRADE	174.37	50	ARCA	Regular, FormT, OddLot	
NXPI	16:53:11	16:53:11	TRADE	174.37	1	ARCA	Regular, FormT, OddLot	
NXPI	16:58:09	16:58:09	TRADE	174.37	1	ARCA	Regular, IntermarketSweep, FormT, OddLot, TradeThroughExempt	

Source of Table Data: algoseek recorded live SIP feed

Example 2: "Invented" Volume

Ticker	Date Time	Vendor	Volume
NXPI	2021/01/25 9:29	algoseek	10
NXPI	2021/01/25 9:29	Bloomberg	10
NXPI	2021/01/25 9:29	Vendor R	2,611

Table: algoseek vs Bloomberg vs Vendor R (ITRM on January 25)

“Market data becomes commodity” is a lie

- **Effective market theory**: Share prices reflect all information.
- Driving/mirroring share price changes are **market microstructure and dynamics**, for example
  - shares traded at bid/ask/mid
  - number of upticks/downticks
  - hidden orders/icebergs
  - retail trader activities
- Live data feeds from exchanges contain **detailed information on quote/trade sources and conditions**
- Change of data needs driven by **Machine Learning**
  - There are large differences between data for machines and data for human
  - algoseek data is designed for machine learning and quantitative trading/research
- While smart people focus on alternative data, algoseek persistently focuses on developing and perfecting market data.

algoseek Comprehensive Data Products

algoseek's deep understanding on quant and algo trading has led to **the most comprehensive and detailed market data products in the financial data industry**. For example:

	algoseek THE MARKET DATA COMPANY	Other Vendors
Number of Data Fields in Minute Bars (for all asset classes)	80	10 - 20
aggressor flag (for Futures Trades) & **aggressor count** (for Futures Minute Bars)	Yes	No
Continuous Bar Every Minute (for Equities Minute Bars) (always show open/close bid/ask)	Yes	No
Choices of Exchanges (for all Equities aggregated data)	Both **All Exchanges and TRF** & **Listed Exchanges Only**	Only All Exchanges and TRF
TAQ & TANQ (for Options)	Both **TANQ** (Trades & Quotes) & **TAQ** (Trades & NBBO Quotes)	Only TANQ

Examples of algoseek TAQ Minute Bar Data Points

CancelSize
VolumeWeightPrice
TimeWeightedBid
TimeWeightedAsk
TradeAtBid
TradeAtBidMid
TradeAtMid
TradeAtMidAsk
TradeAtAsk
TradeAtCrossOrLocked
FinraVolume
FinraVolumeWeightPrice
UptickVolume
DowntickVolume
RepeatUptickVolume
RepeatDowntickVolume
UnknownVolume
TradeToMidVolWeight
TradeToMidVolWeightRelative

algoseek data provides in-depth information on market microstructure and dynamics

algoseek
THE MARKET DATA COMPANY


algoseek
THE MARKET DATA COMPANY