TICK DATA

High Frequency Data Filtering

A review of the issues associated with maintaining and cleaning a high frequency financial database

THOMAS NEAL FALKENBERRY, CFA

As seen on: Wall Street & Technology online

High Frequency Data Filtering

A review of the issues associated with maintaining and cleaning a high frequency financial database

Thomas Neal Falkenberry, CFA

Every day millions of data points flow out of the global financial markets driving investor and trader decision logic. These data points, or ticks,™ represent the basic building blocks of analysis. Unfortunately, the data is too often transmitted with erroneous prices that render pre-filtered data unusable.

The importance of clean data, and hence an emphasis on filtering bad data, has risen in recent years. Advances in technology (bandwidth, computing power, and storage) have made analysis of the large datasets associated with higher frequency data more accessible to market participants. In response, the academic and professional community has made rapid advances in the fields of trading, microstructure theory, arbitrage, option pricing, and risk management to name a few. We refer readers to Lequeux (1999)[i] for an excellent overview of various subjects of high frequency research.

In turn, the increased usage of high frequency data has created the need for electronic execution platforms to act on the higher frequency of trade decisions. By electronic execution, we do not refer to the process of typing order specifications into a Web site and having the order electronically transmitted. We refer to the fully-automated process of electronically receiving data, processing that data through decision logic, generating orders, communicating those orders electronically, and finally, receiving confirmation of transactions. A bad tick into the system means a possible bad order out of the system. The cost of exiting a trade generated on a bad tick becomes a new source of system slippage and a potential huge source of risk via duplicate or unexpected orders.

Estimates for the frequency of bad ticks vary. Dacorogna *et al.* (1995)[ii] estimated that error rates on forex quote data are between 0.11% and 0.81%. Lundin *et al.* (1999)[iii] describe the use of filters in preprocessing forex, stock index, and implied forward interest rate returns whereby 2%–3% of all data points were identified as false outliers.

This paper will describe the issues associated with maintaining and cleaning a high frequency financial database. We will attempt to identify the problem, its origins, properties, and solutions. We will also outline the filters developed by Tick Data, Inc. to address the problem, although the outline is intentionally general. This paper will make frequent use of charts and tables to illustrate key points. These charts and tables include data provided from multiple sources, each of which is highly reputable. The errant data points illustrated in this paper are structural to the market information process and do not reflect problems, outages, or the lack of quality control on the part of any vendor.

Thomas Neal Falkenberry, CFA, is President of Tick Data, Inc. He is also the founder of Autumn Wind Asset Management, an SEC-registered investment advisory firm, and the General Partner to Autumn Wind Capital Partners, L.P., a commodity pool operator. He may be reached at (703) 757-1370 or tnf@tickdata.com.

I. The Problem

Intraday data, also referred to interchangeably as tick data and high frequency data, is characterized by issues that relate to both the structure of the market information process as well as to statistical properties of data itself.

At a basic level the problem is characterized by size. Microsoft (MSFT) has averaged 90,000 ticks per day over the past twelve months. That equates to 22.6 million data points for a single year. While the number of stocks with this high level of tick count is limited, the median stock in the Russell 3000 produces approximately 2,100 ticks per day or 530,000 per year. A reasonable research or buy list of 500 stocks, each with three to five years of data, can exceed two billion data points. Data storage requirements can easily reach several hundred gigabytes after storing date, time, and volume for each tick.

While advances in databases, database programming, and computing power have made the size issue easier to manage, the statistical characteristics of high frequency data leave plenty of challenges. Specifically, problems arise due to:

- The asynchronous nature of tick data.
- The myriad of possible error types, including isolated bad ticks, multiple bad ticks in succession, decimal errors, transposition errors, and the loss of the decimal portion of a number.
- The treatment of time.
- Differences in tick frequency across securities.
- Intraday seasonal patterns in tick frequency.
- Bid-ask bounce.
- The inability to explain the cause of errant data.

Yet, perhaps the most difficult aspect of cleaning intraday data is the inability to universally define what is "unclean." You know it when you see it, but not everyone sees the same thing. There are obvious outliers, such as decimal errors, and there are borderline errors, such as losing the fractional portion of a number or a trade reported thirty seconds out of sequence. The removal of obvious outliers is a relatively easy problem to solve. The complexity lies in the handling of borderline, or marginal, errors.

The filtering of marginal errors involves a tradeoff. Filter data too loosely and you still have unusable data for testing. Filter data too tightly and you increase the possibility that you overscrub it, thereby taking reality out of the data and changing its statistical properties. **Overscrubbing data is a serious form of risk.** Models that have been developed on overscrubbed data are likely to find real-time trading a chaotic experience. Entry and exit logic based on stop and limit orders will be routinely triggered from real-time data that demonstrates considerably greater volatility than that experienced during simulation. In Dunis *et al.* (1998)[iv] a methodology for tick filtering is described whereby the authors state, *"cleaning and filtering of an archived database will typically be far more rigorous than what can feasibly be achieved for incoming real-time data."* We reject this concept for the reason sited above. Treating data differently in real time versus historical simulation can be risky.

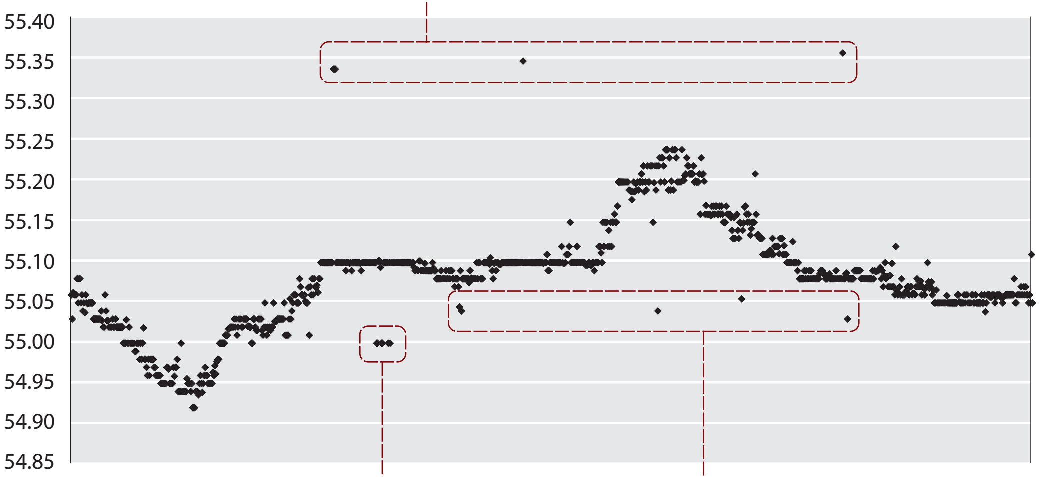
Defining marginal errors is the crux in the tradeoff between underscrubbing and overscrubing data. In our opinion, these errors are a function of the base data unit (tick, 1-minute, 60-minute, etc.) employed by the trader. What is a bad tick to a tick-based trader may be insignificant to a trader using 60-minute bars. That is not to say that the 60-minute trader cannot or should not filter data to the same degree as the tick trader, but the decision to do so may unnecessarily add to the level of sophistication required by the filter(s). This unconventional idea, that **error definition is unique to the trader and hence, that there is no single correct scrubbed time series applicable to all traders**, has evolved through our work with high frequency data and traders over the past eighteen years. We believe it is more important to match the properties of historical and real-time data than it is to have “perfect” historical data and “imperfect” real-time data.

The primary objective in developing a set of tick filters is to manage the overscrub/underscrub tradeoff in such a fashion as to produce a time series that removes false outliers in the trader’s base unit of analysis that can support historical backtesting without removing real-time properties of the data.

Graphical Representation of the Problem

The chart below contains tick data for MSFT for May 2, 2002, from 9:40–9:43 and is a fair representation of the general problem.

Microsoft (MSFT) | May 2, 2002, 9:40–9:43 am



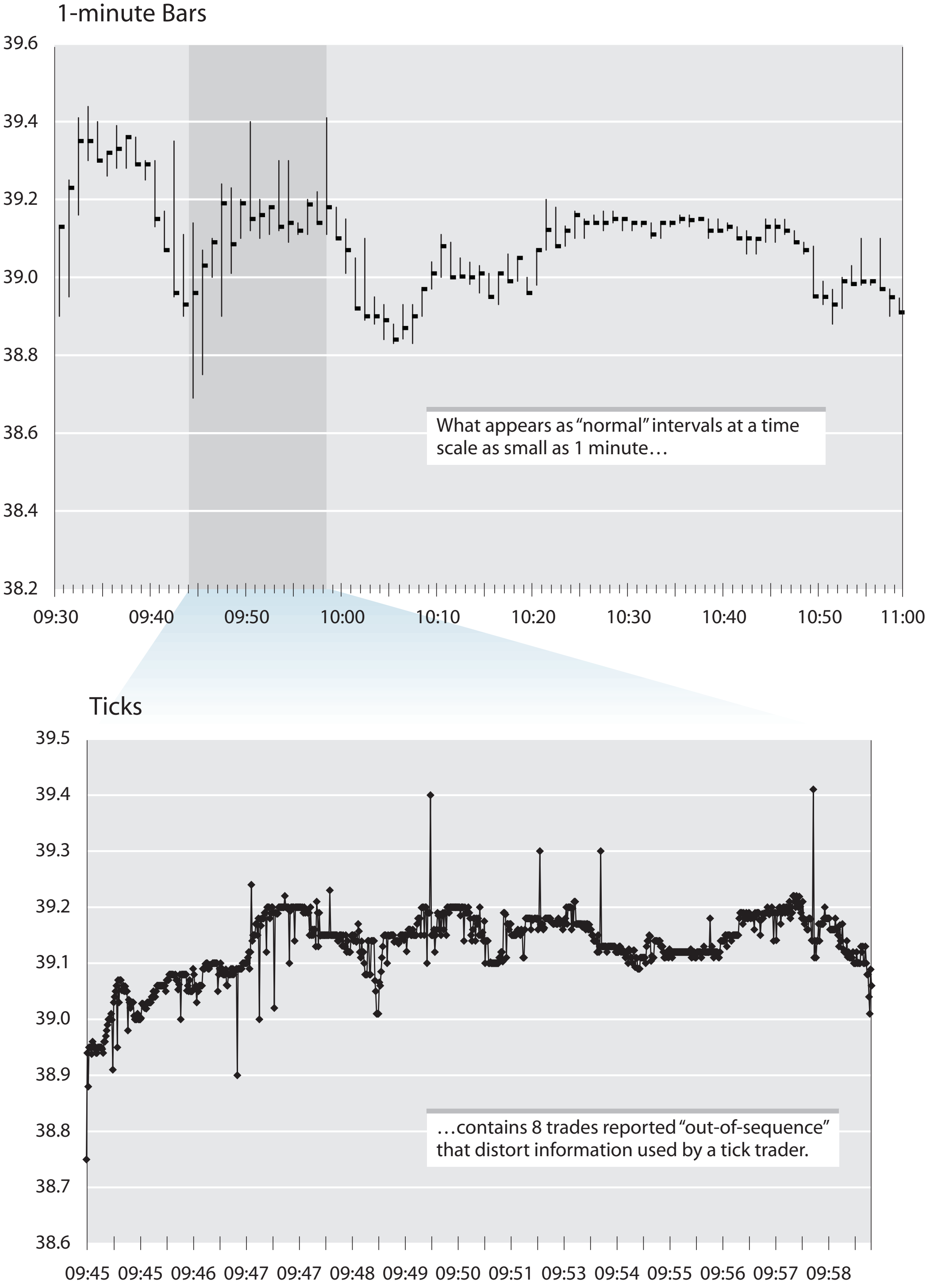
Data are easily recognized as isolated bad ticks. The prints are 0.25–0.35 points off the market and are followed by prices that return to prior levels.

Data points represent 3 bad ticks in succession. Interestingly, the bad ticks lie at the value 55.00. Most likely, these ticks appear as bad because the fractional portion of the price was "lost."

Data points represent ticks that are questionable. These prints are bad to some traders and irrelevant to others. This is where the difficulty in filtering data lies, as this is where the trade-off between underscrubbing/ overscrubbing is determined.

The MSFT chart may appear to illustrate problems that are "non-issues" to users of slower, e.g. 45-minute or daily, data. Users of slower data are affected by bad high frequency data, but the problems are simply less obvious than they are to a user of higher frequency data. For example:

Costco (COST) | May 30, 2002, 9:30–11:00 am



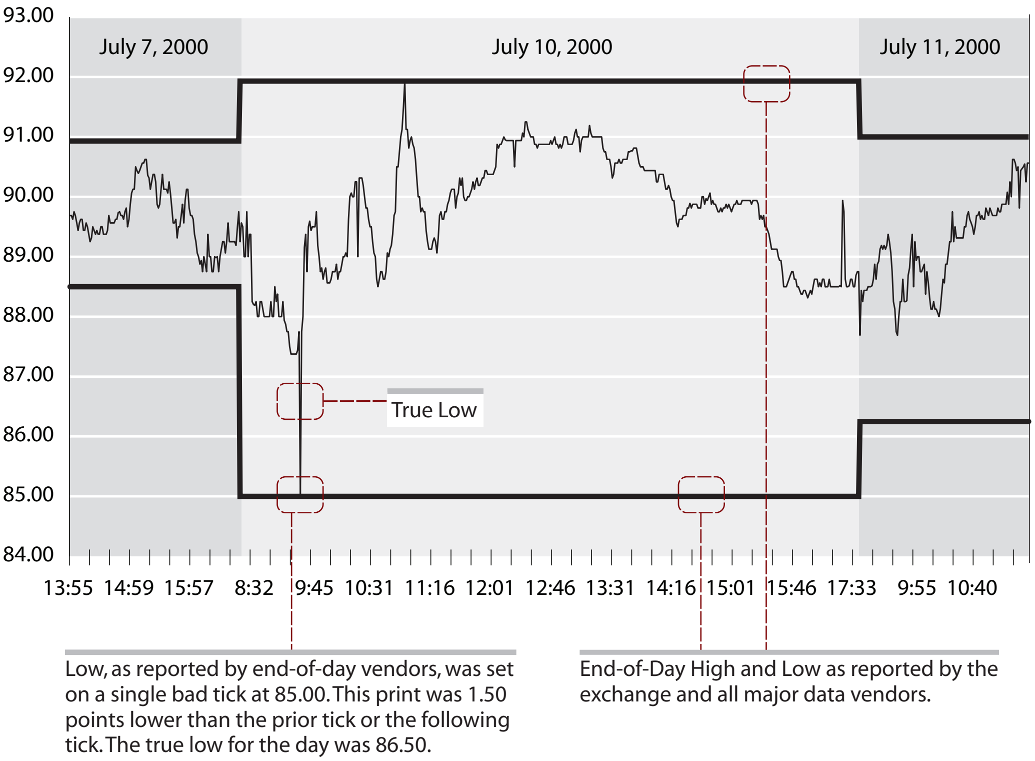
1-minute Bars

What appears as "normal" intervals at a time scale as small as 1 minute…

Ticks

…contains 8 trades reported "out-of-sequence" that distort information used by a tick trader.

Applied Materials (AMAT) | July 7, 2000–July 11, 2000



Low, as reported by end-of-day vendors, was set on a single bad tick at 85.00. This print was 1.50 points lower than the prior tick or the following tick. The true low for the day was 86.50.

End-of-Day High and Low as reported by the exchange and all major data vendors.

We estimate that the high or low of daily data, or any intraday bar, is set on a bad tick far more frequently than is currently perceived. However, most users of daily data have no means by which to view the tick activity associated with setting these daily extremes.

This problem reflects the possible fractal nature of security prices. What appears to be “clean” daily data actually contains “unclean” 45-minute bars. Drill down into those “clean” 45-minute bars and you will find 1-minute bars that are unusable to a trader with a shorter base unit of analysis. Likewise, there are bad ticks within those 1-minute bars that go unrecognized to all but tick traders. What looks acceptable to one trader looks bad to a trader in a shorter time scale. This is true for any time scale greater than tick level. **Data must be cleaned at its finest granularity.**

While filtering must take place at the tick level, filter parameters should remain a function of the trader’s base unit of analysis. Filters to clean ticks to the level required by the tick trader are computationally more difficult than filters to remove outliers in 45-minute time space. **Again, there is no single correct scrubbed time series. The objective of scrubbing data is to remove aberrant data in the traders base unit of analysis in such fashion that does not change the statistical properties of the data vis-à-vis a real-time datafeed.**

II. Why Bad Data Exists

The source of errant data points is difficult to assess. There are many. Yet the root of the problem can be traced to the speed and volume at which the data is generated and human intervention between the point of trade and data transmission. A basic review of the mechanics of order execution is useful in understanding the problem.

Open Outcry or Auction Markets

In open outcry or auction markets, trades are generated on physical exchanges by participants voicing, rather shouting, bids and asks. Trades are recorded by pit reporters located on the floor near the parties agreeing on the terms of a trade. For example, in the US Treasury Bond pit at the CBOT, there are five market reporters stationed in a tower located above the pit and an additional three reporters located on the floor in the center of the pit. These reporters are trained to understand the various hand signals and verbal queues that confirm trade, bid, and ask prices. Reporters enter prices into handheld devices and route them to servers that in turn release them to data vendors. Software running on the handheld devices reduces, but does not eliminate, the possibility of multiple reporters entering the same trade. A similar structure exists with the NYSE whereby trading assistants record the transactions of the specialists.

It is not hard to see how data errors can emerge from this process, particularly in fast markets. In fact it is remarkable that the process works as well as it does. Humans can only accurately type so fast and sequence so many trades.

Electronic Markets

Electronic trading represents a logical technological progression from open outcry. In these markets, there are no physical exchanges. Buyer and seller orders are matched electronically. Pit reporters are replaced with servers. Data is collected electronically in sequence and released to data vendors.

Is data from electronic markets cleaner than data from open outcry or auction markets? Logic would support the argument, but reality does not. For example, compare a representative NASDAQ symbol (electronic) to one from the NYSE (auction). Compare the front-month S&P500 futures contract traded in the CME pit to the DAX contract traded electronically on EUREX. The first impression is that electronically traded symbols actually experiences higher error rates. The reason for this, we believe, has nothing to do with the method by which trading occurs, but rather the volume of trading itself. For example, for May 2002, the electronic DAX contract averaged 26,000 ticks per day versus 3,300 for the pit traded S&P contract. The largest NYSE company, GE, averaged 22,000 ticks per day versus 90,000 for the electronically traded MSFT. Tick frequency is a better predictor of error rates than whether the item is traded electronically or in open outcry.

The most common causes of bad data are:

- Human error in the face of high volume. This broad genre includes decimal errors, transposition errors, losing the fractional portion of a number, and simply bad typing.
- Bad data comes from processes inherent to trading. Various scenarios can arise whereby trades are reported out-of-sequence, sold bunched, cancelled, cancel and replaced, and reported in error just to name a few. Refer to https://www.nasdaqtrader.com/easp/reportsource.htm under Report Keys and Samples for a full listing of NASDAQ trades codes. Similar codes exist for all major exchanges.
- Bad data comes from multiple markets simultaneously trading the same security. The following table shows floor trades and regional trades for AOL from January 12, 1999:

Symbol	Shares	Price	Time
AOL	4500	163 ½	9.38
AOL	7700	163	9.38
AOL	26300	162 ½	9.38
AOL	1400	162 ½	9.38
AOL	25800	162	9.38
AOL	30000	162	9.38
AOL	9400	162	9.38
AOL	5500	162 1/16	9.38

3 point difference between floor trading and regional trading will cause an unusually large range for the 9:38 interval.

Symbol	Exchange	Shares	Price	Time
AOL	BO	300	165	9.38
AOL	BO	300	165	9.38
AOL	BO	200	165	9.38
AOL	BO	200	165	9.38
AOL	BO	100	165	9.38
AOL	MW	200	165	9.38
AOL	BO	1500	165	9.38
AOL	BO	400	165	9.38
AOL	BO	1000	165 1/16	9.38
AOL	BO	100	165	9.38
AOL	BO	600	165	9.38
AOL	BO	200	165	9.38
AOL	BO	200	165	9.38
AOL	BO	100	165	9.38
AOL	NW	200	165	9.38

In summary, bad data emerges from the asynchronous and voluminous nature of financial data. The system goes from an "off" state, overwhelms recording mechanisms during the flurry around opening trading and news events, settles down, and returns to an "off" state. Global events then arise that serve as the source for the following opening's activity. Simultaneously, there are thousands of investment professionals who occasionally make trading errors that must be canceled, replaced, and corrected. There are human limitations and errors in the recording process. There are technological glitches and outages thrown in as well. It is a credit to the professionals involved in the process that is runs as well as it does.

III. Specific Properties of Equity Tick Data

One property of equity data that adds to the complexity of the filtering problem is the dramatically different activity levels of various issues. The following table lists market capitalization, total ticks, and total volume for various constituents of the Russell 3000 for May 23, 2002.

Russell 3000 Constituents | May 23, 2002

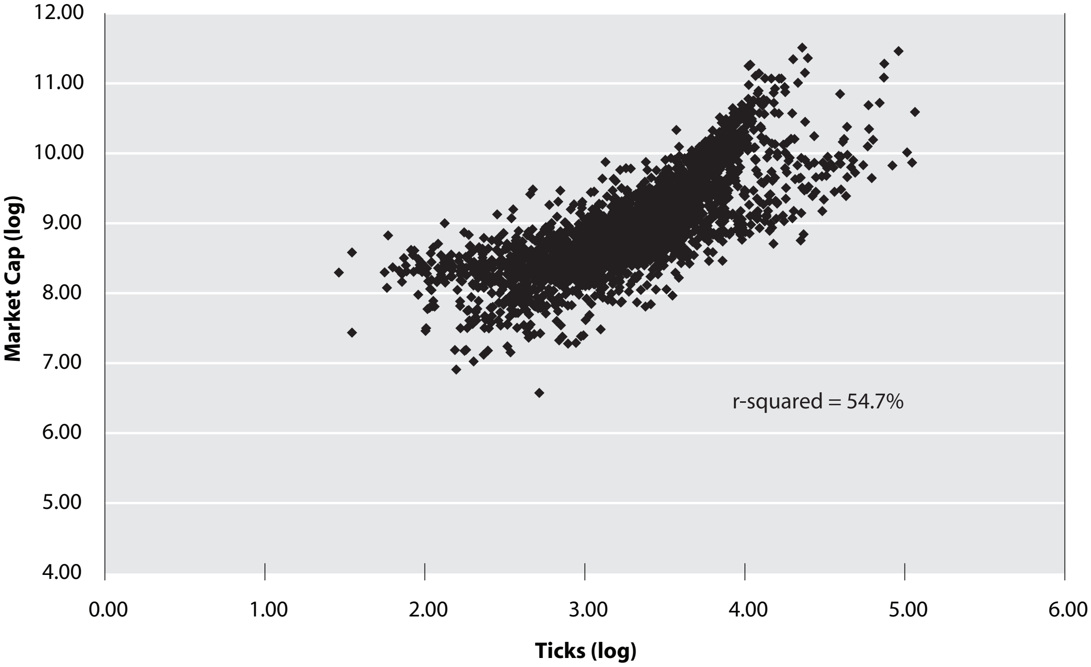
Rank by Cap	Company Name	Market Cap	Total Ticks	Total Volume
1	GENERAL ELECTRIC CO	323,971,300,000	22,852	11,250,400
2	MICROSOFT CORP	288,427,500,000	91,439	18,007,800
3	CITIGROUP INC	230,691,500,000	24,767	6,112,500
4	PFIZER INC	221,624,900,000	20,053	7,499,800
5	INTEL CORP	191,620,800,000	74,424	26,050,100
6	JOHNSON & JOHNSON	184,888,000,000	10,825	5,009,000
7	AMERICAN INTERNATIONAL GROUP	177,154,800,000	10,539	3,110,400
8	INTL BUSINESS MACHINES CORP	142,292,300,000	23,742	3,233,000
9	COCA-COLA CO/THE	139,585,100,000	12,261	3,070,600
10	MERCK & CO. INC.	127,959,600,000	11,672	3,493,500
11	CISCO SYSTEMS INC	121,060,400,000	74,048	41,703,600
12	PHILIP MORRIS COMPANIES INC	117,910,800,000	13,358	3,417,200
	…	…	…	…
77	HOUSEHOLD INTERNATIONAL INC	24,227,440,000	12,543	1,930,500
78	QUALCOMM INC	23,867,670,000	43,484	6,018,100
79	TENET HEALTHCARE CORPORATION	22,999,650,000	7,852	963,500
80	PHILLIPS PETROLEUM CO	22,689,470,000	12,678	1,217,700
81	METLIFE INC	22,319,830,000	7,530	1,707,900
82	SUN MICROSYSTEMS INC	22,269,410,000	59,673	57,341,300
83	ILLINOIS TOOL WORKS	22,119,120,000	7,739	1,065,200
	…	…	…	…
228	FORTUNE BRANDS INC	8,191,684,000	7,170	588,100
229	AMERISOURCEBERGEN CORP	8,189,010,000	5,688	416,600
230	NORFOLK SOUTHERN CORP	8,107,163,000	6,817	978,100
231	M & T BANK CORP	8,102,100,000	3,651	145,300
232	SUNGARD DATA SYSTEMS	8,069,937,000	4,846	610,100
233	EQUITY RESIDENTIAL	8,067,264,000	4,531	692,600
234	FISERV INC	8,062,585,000	23,793	1,380,500
235	JOHNSON CONTROLS INC	8,061,801,000	3,655	187,500
236	AMSOUTH BANCORPORATION	8,049,972,000	5,526	487,600
	…	…	…	…
1,530	ELCOR CORP	500,336,900	1,333	42,900
1,531	INTL MULTIFOODS CORP	499,118,800	941	17,700
1,532	CELL GENESYS INC	498,977,100	3,596	254,500
1,533	INTEGRA LIFESCIENCES HOLDING	498,756,000	1,292	108,300
1,534	GYMBOREE CORP	498,241,100	4,001	301,400

Tick activity differs greatly across securities. This poses a significant problem in developing a filter as it introduces time as a variable. For example, with 91,000 ticks per day, MSFT averages a tick every .333 seconds whereas Illinois Toolworks, still in the mid cap universe, averages one tick every three seconds. Jump to small cap issues and ticks may roll in every three to five minutes. A filter must be able to handle the difference in time elapsing between ticks as it directly influences the amount of price movement that can occur prior to a tick being suspected as aberrant. For example, a tick $0.20 off the prior tick may be suspect for an issue generating multiple ticks per second, but may be acceptable for a mid cap issue generating a tick every four minutes. It is far more likely that new information entering the price discovery process leads to a $0.20 price change over four minutes than it is over one-fourth of a second.

Traders generally focus their trading on liquid issues. Market cap is generally used as a proxy to identify candidates. This is often done because of an assumed relationship between market cap and liquidity and the ease of obtaining market cap statistics. A more thorough definition of "tradable stocks" should read: *"Liquid (volume) issues that are actively traded (tick count) where slippage is likely to be minimal (volume (if trading size is large) and tick count)."* The following two charts plot these relationships.

Relationship between Ticks and Market Capitalization

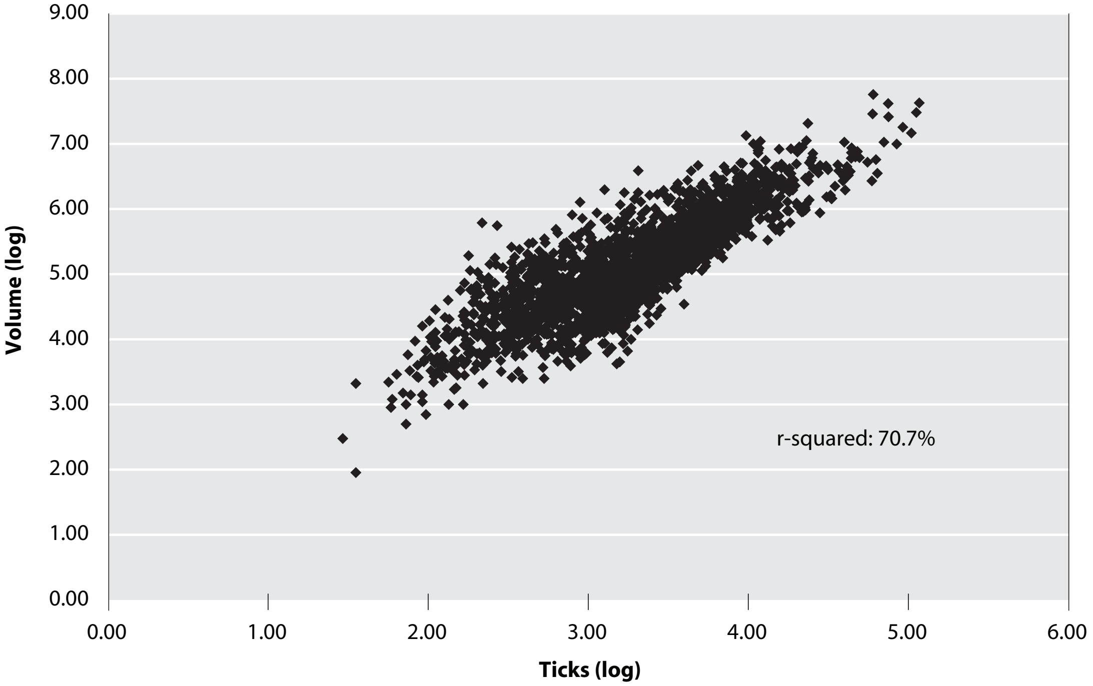
Russell 3000 Constituents | May 23, 2002



The relationship between tick activity and market cap is high. The lack of a tighter relationship is due to the disproportionately high tick activity of "usual suspect" stocks in a few industry groups, namely technology and biotechnology. Mid cap names such as Biogen (BGEN), Qlogic (QLGC), Emulex (EMLX), and NVIDEA (NVDA) have two to three times the tick activity of General Electric (GE) but only 1.5% of its market cap.

Relationship between Ticks and Volume

Russell 3000 Constituents | May 23, 2002



The relationship between volume and tick count is tighter. Deviation is again attributable to a few technology and biotechnology issues.

The implication of these relationships is too highlight the complexity in developing a "one size fits all" filter. Equity issues are not homogeneous time series. An issue that averages a tick every four minutes may be just as difficult to filter as an issue with multiple ticks per second. The later requires speed of calculation, the former a tolerance for greater price movement due to the increased time passing between ticks. **A filter, or set of filters, requires parameters that can adapt to tick frequency in order to address the effect of time. This, in turn, implies that filters must adapt to volatility of price movement as well.**

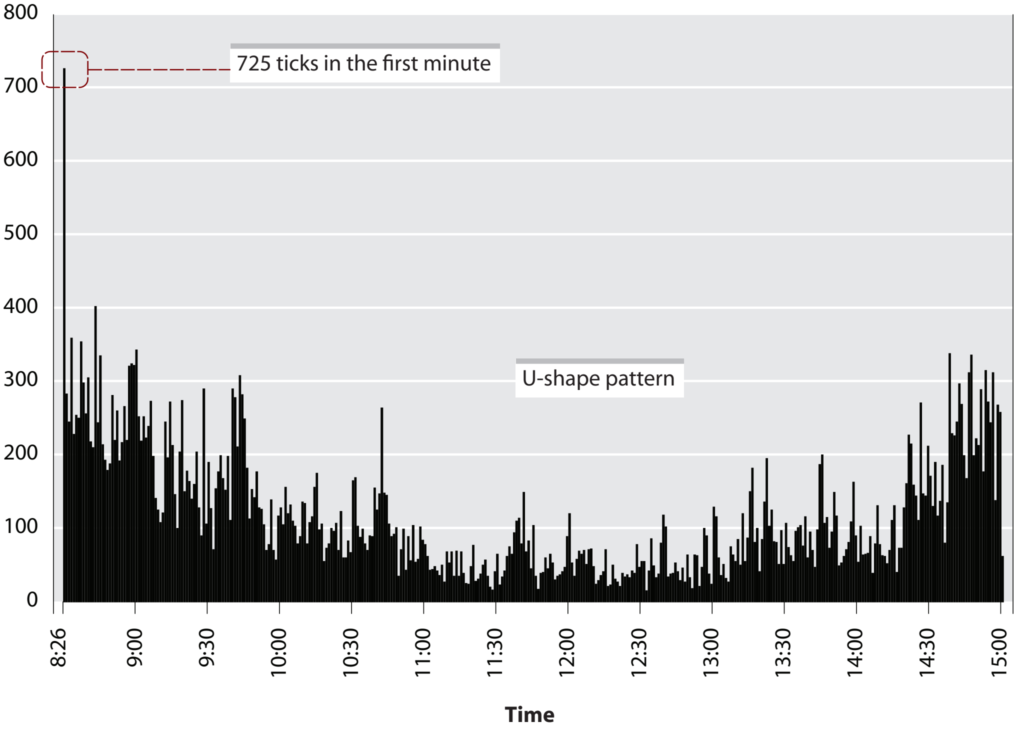
Intraday Tick Patterns: Ticks per Minute

In addition to equities differing by tick frequency, issues also demonstrate distinct intraday seasonal patterns. We believe there are three general groupings:

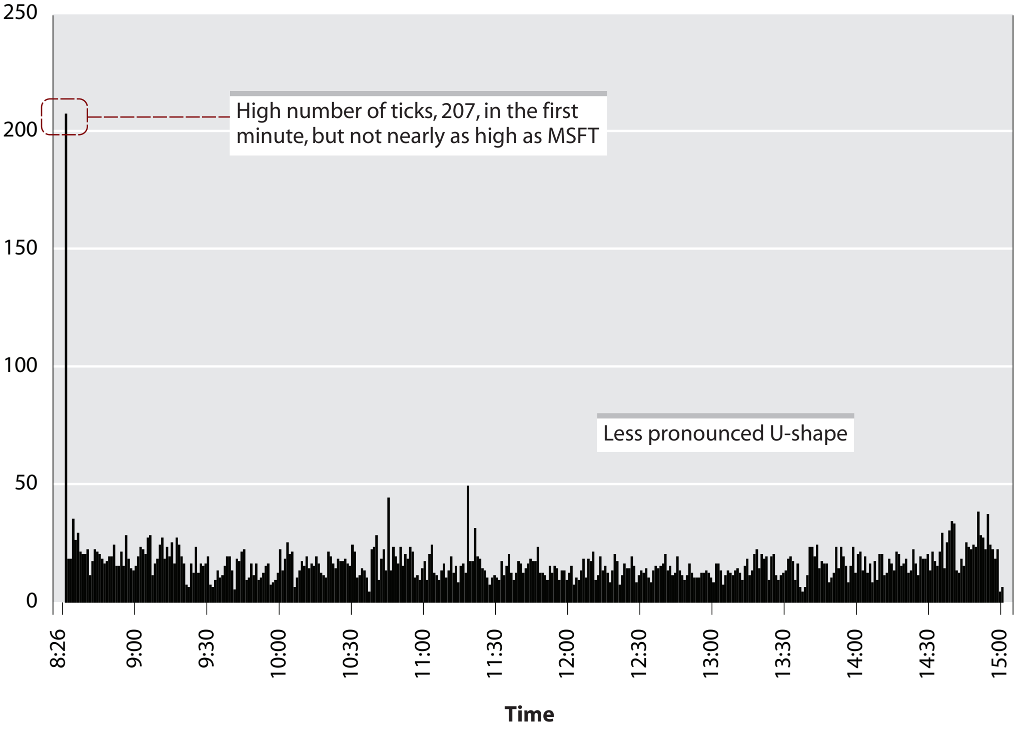
Group I	High volume NASDAQ issues,
Group II	Large cap NYSE and large cap NASDAQ issues not included in Group #1, and
Group III	Small and mid cap issues on all exchanges.

The following charts display average ticks per minute for the month of April 2002.

Group I (MSFT) | Representative of high volume NASDAQ issues (INTC, CSCO)



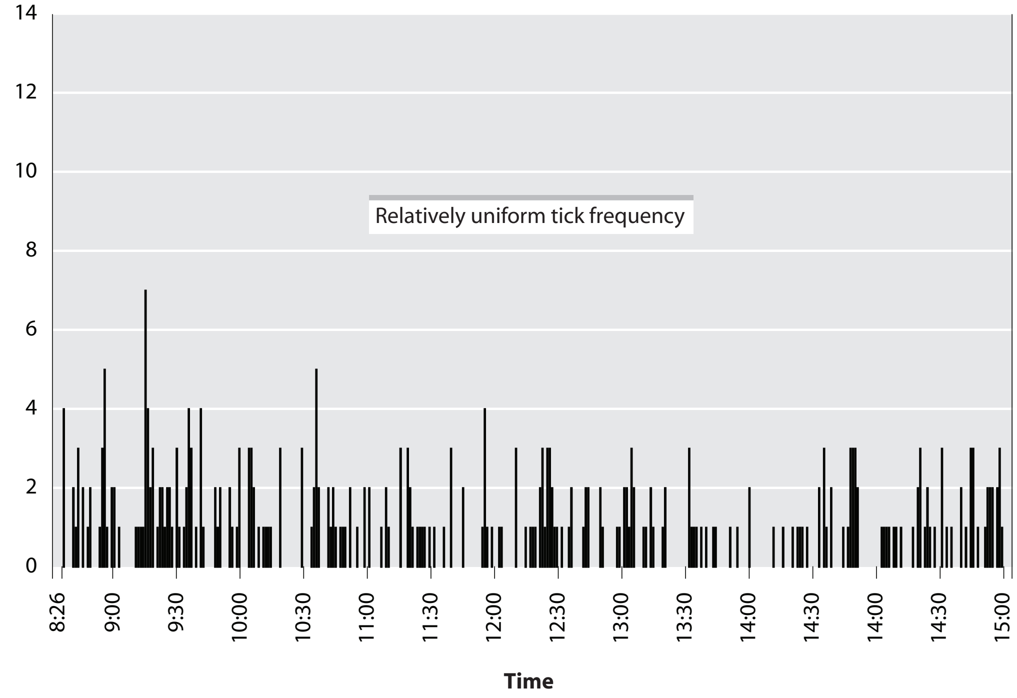
Group II (PFE) | Representative of large cap NYSE issues (GE, TYC, JNJ)



High number of ticks, 207, in the first minute, but not nearly as high as MSFT

Less pronounced U-shape

Group III (IOM) | Representative of small and mid cap issues



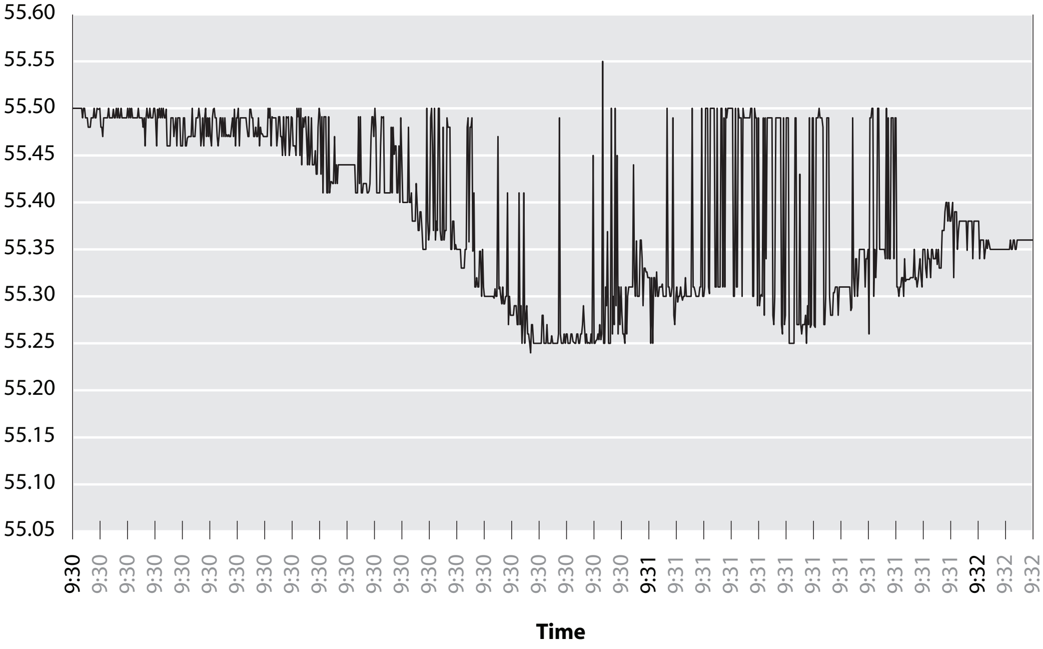
Relatively uniform tick frequency

Intraday seasonal patterns show an enormous number of ticks in the opening minutes for Group I and Group II stocks. Tick frequency then demonstrates the well-documented U shape,[v] reaching a lull at midday prior to increasing towards the close. This pattern is less pronounced for listed issues. Smaller cap issues tend to demonstrate consistent tick frequencies throughout the day.

The implication of differing seasonal patterns on tick filtering is a variation on the theme described earlier with tick frequency. As the amount of time passing between ticks varies, whether due to tick frequency differences between issues or the time of day effect within an issue, a filter must adapt to the volatility of price changes that is a function of tick frequency.

As stated, the opening few minutes is characterized by high tick count and high volatility. The following chart is a representative case.

Microsoft (MSFT) | May 20, 2002, 9:30–9:32 am



“Flagged” out of sequence trades have been removed from the chart above. Prices swing wildly from 55.25 to 55.50. Are there “unflagged” out of sequence trades? Probably, as there were 3,352 trades in the 180 seconds plotted on the graph.

Lastly, equity issues differ dramatically in price. As of May, 2002, the highest priced stock in the Russell 3000 was at a level of $715.25. The lowest priced was $0.22. Obviously, absolute price changes from tick to tick cannot serve to flag bad data and would certainly not scale across securities.

As can be seen, tick filters must be dynamic and adaptive. They must handle:

- Tick frequencies that differ across securities.
- Tick frequencies that differ intraday within the same security.
- Price levels that differ across securities.

IV. Solutions

To review, a filter should

1. Create a time series for historical research that eliminates outliers in the trader's base unit of analysis without introducing concepts and techniques that cannot be applied in real-time.
2. Not change the statistical properties of data relative to that which will be used in real-time.
3. Not introduce excessive delay due to computation time or the need for excessive confirming data points, i.e. a suspected bad tick at time *t* being confirmed by future prices generated at time *t*+1, *t*+2, etc.
4. Be adaptive across securities with different tick frequency profiles.
5. Be adaptive across securities with different price levels.

There are two general approaches to data filtering:

1. "Search and Replace/Delete" bad ticks in the original time series.
   a. Once a bad price is identified, should you delete the tick, replace it with the last known good value, or replace it with another value?
   b. If deleted, do you assign the tick's volume to the previous tick or eliminate volume altogether?
2. Capture "basic price activity" in a separate synthetic time series. Specifically, create a time series consisting of some close representation of the data. For example, a moving average of price captures basic price action (not recommended).

A number of high frequency filters have been published over the past several years. They range from simple moving averages of price to complex heuristic algorithms. Most, in our opinion, are based on valid statistical concepts, but fail to make an explicit connection between identically filtering historical and real-time data. Many of the filters simply are not executable in real-time.

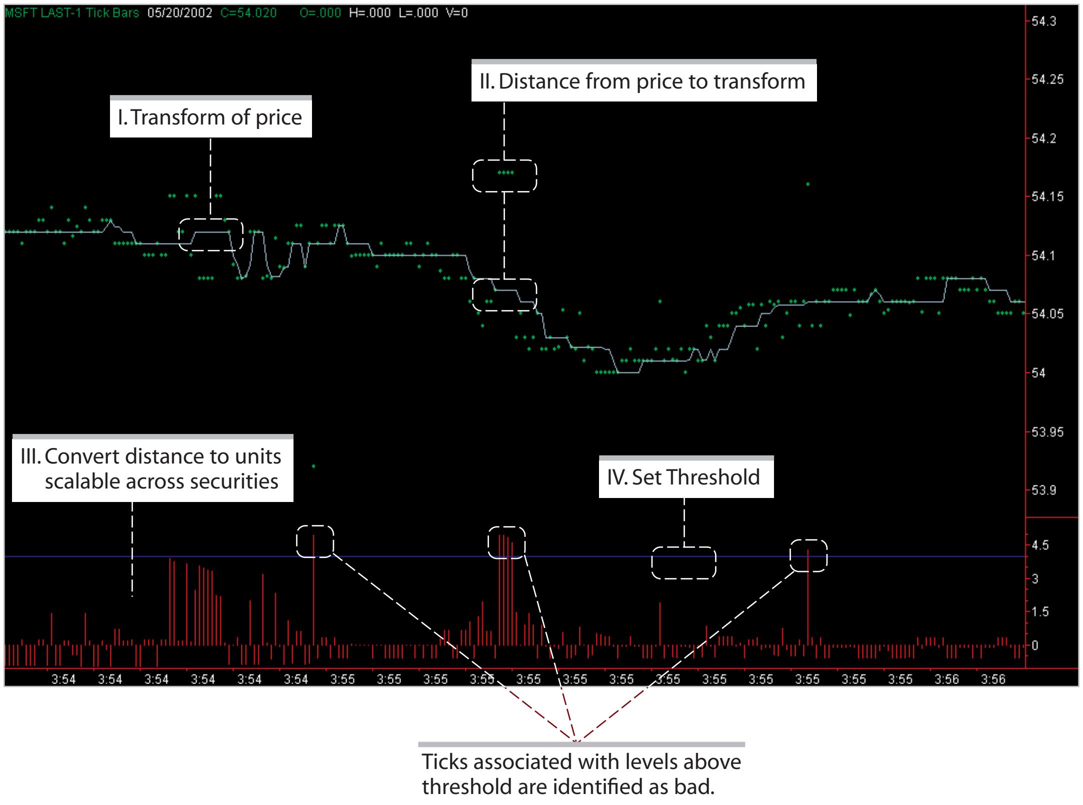
The Tick Data, Inc. Filtering Process

As stated, the purpose of this paper is to overview the subject of high frequency data filtering and briefly describe the methodologies employed by Tick Data, Inc. This paper is not intended to fully disclose the filtering process. Our process is fully disclosed to clients, who believe our methodologies offer them a competitive advantage. Hence, full disclosure is not appropriate in a paper intended as an overview on the subject.

The core premise behind the Tick Data, Inc. filter is to modify as few data points as is necessary to filter historical data in such fashion as to be useful for historical testing and representative of that which will be experienced in real time. As such, we utilize the "Search and Modify" approach described above. We modify errant ticks rather than discard them as we wish to maintain the volume associated with a tick even if its price is bad.

The basis of the filter is a moving transform of price. The number of data points used in calculating the transform is a function of tick frequency. This is the first step in adapting the filter to the unique activity levels of various issues. Next, we measure each tick's distance from this moving transform and convert that difference into units that scale across securities. This makes the filter adaptable to securities with different price scales. Ticks that exceed a user-defined threshold are deemed bad. Allowing the threshold to be defined by the user enables the filter to adapt to the base unit of analysis of the trader and manage the overscrub/underscrub tradeoff. Lastly, ticks that are deem bad are replaced with the value of the transform and assigned the volume of the bad tick.

For example:

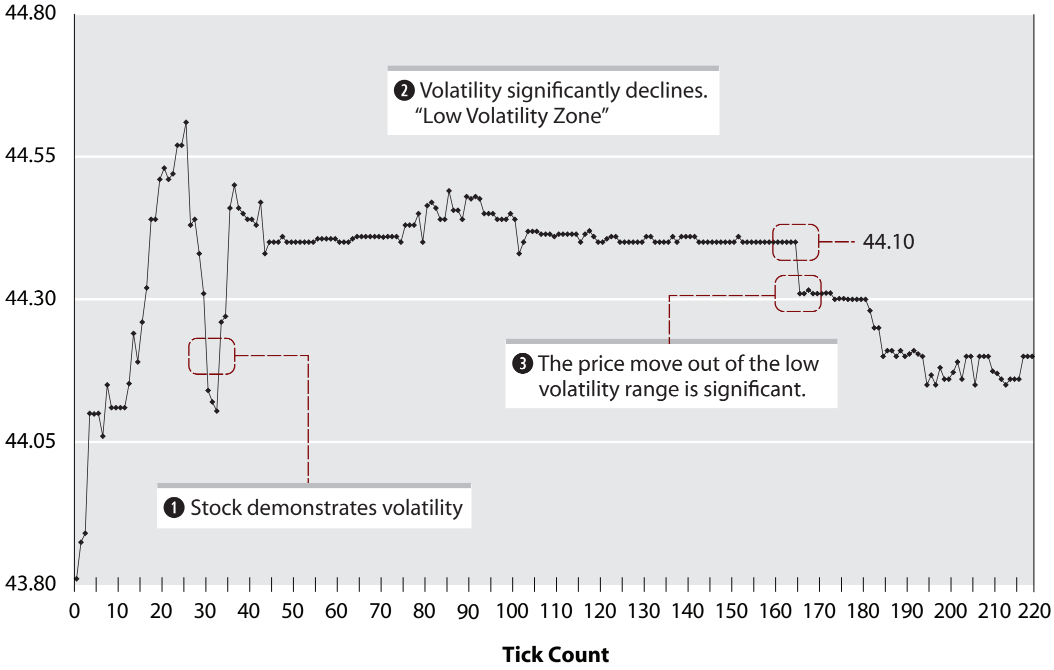


The data can be filtered finer or coarser based upon selection of the threshold. Generally, tick-based traders will seek a lower threshold than will a trader employing 15-minute bar data. A trader using 45-minute data may choose to bypass the use of a threshold altogether and base trading on the transform directly. The later approach seeks to capture "basic price activity" in a separate synthetic time series. The former is the "Search and Replace" method. Our preference is to search and replace, leaving as many valid ticks undisturbed as possible. We note, however, that the two methodologies differ very little at longer time frames. The methodology described above, regardless of threshold selection, can be employed in real-time.

V. Filter Limitations

a. Edge Effects – Low Volatility Zones

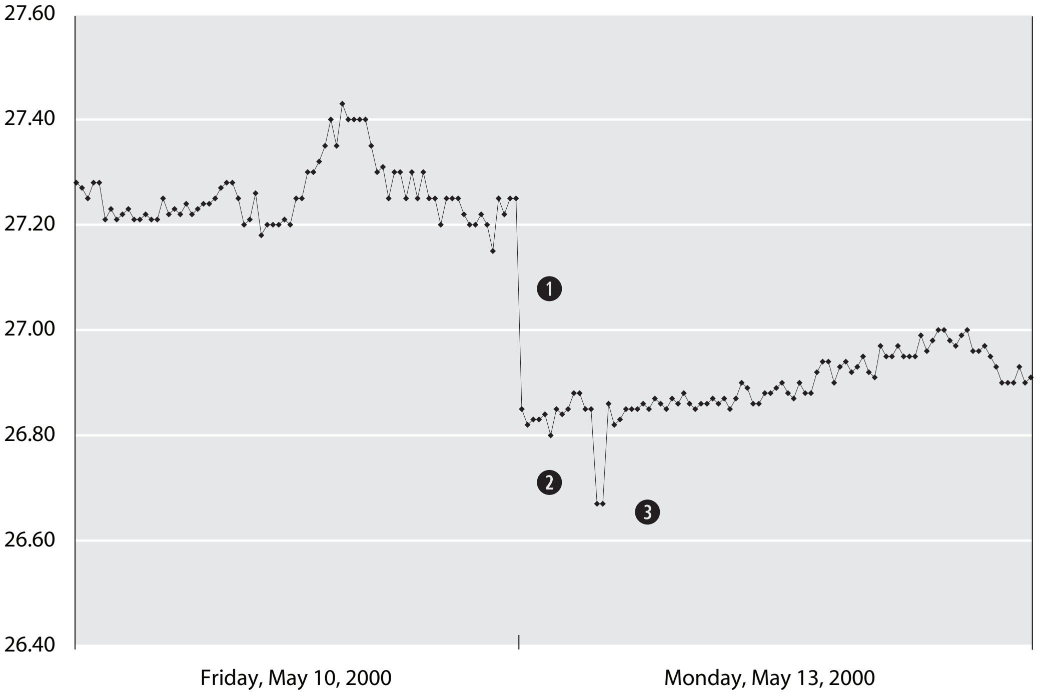
MXIM | June 13, 2002, 13:23–13:52



The last tick of 44.40 in the area of low volatility is followed by prints at ❸ of 44.31, 44.31, and 44.35, respectively. Our scrubbers would identify the first 44.31 print as bad and change the value of the print to the value of the transform, 44.40. The second 44.31 print, and all subsequent prints shown on the chart would remain unfiltered. We believe sacrificing the first print coming out of a low volatility zone is an acceptable tradeoff given the range of data problems correctly identified and repaired by the filters.

b. Edge Effects – First n *Ticks of the Day*

NYMEX CRUDE OIL (CLN2 Contract) | May 10 and May 13, 2002



❶ The overnight or weekend gap effect renders the first tick of the day unfilterable.
❷ The effect of the gap also distorts the short-term volatility measures we use to judge a tick's validity. To counter this effect, we reinitialize all transforms daily. This means we cannot scrub the first few ticks in a trading session. The risk is that the prints identified by ❸ occur prior to the transforms initializing thereby leaving them unfiltered. Given the volatility and tick frequency of the market on the morning of May 13, our filters would have been initialed by the fifth tick of the session, or approximately 23 seconds after the market opened.

*c. "*n*" Bad Ticks in Succession*

There is a limit to the number of bad ticks in succession that we can filter. At a minimum, we can always filter two bad ticks in succession. At a maximum, we can filter nine. The value between these extremes is determined by the tick frequency of the issue at the time of the tick being filtered. If an exchange's reporting mechanism goes awry and reports fifty ticks in succession with decimal errors, we will not be able to filter the problem. Such events are rare, if seen at all.

VI. Conclusion

The use of high frequency data appears unavoidable. The ability to understand market microstructure is too promising and too possible with recent technological advances to ignore. Yet, the data that underlies this research is bulky, unclean, and difficult to manage. In addition, it contains problems that traditional statistical approaches are not designed to handle. In fact, the problems themselves are difficult to define from trader to trader.

The practical costs of not considering the issues addresses in the paper are:

1. The validity of system research. Has aberrant data in the test set or out-of-sample set had an impact on research results?
2. The acceptance of false positive research results, i.e. accepting models based on over-scrubbed historical data that fail to recognize the properties of real-time data. What appears valid in the lab may fail in reality.
3. Users of electronic execution platforms face a new form of system slippage in exiting erroneous trades entered on stop orders due to poor tick filtering.
4. Developing filters that fail to recognize the unique tick-level properties of different equity issues and time of day effect can lead to overscrubbing/underscrubbing data sets.
5. Maintaining a belief that there is a single correct and perfect time series fails to recognize the complexity of the problem.

Additional advances in technology are certain to elevate these issues to greater priority as traders continue to seek competitive advantage through the use of higher frequency data. While the problems are complex there are solutions.

Thomas Neal Falkenberry, CFA, is President of Tick Data, Inc. He is also the founder of Autumn Wind Asset Management, an SEC-registered investment advisory firm, and the General Partner to Autumn Wind Capital Partners, L.P., a commodity pool operator. He may be reached at (703) 757-1370 or tnf@tickdata.com.

Notes

i. Lequeux Pierre (ed.), 1999, *Financial Markets Tick by Tick: Insight in Financial Market Microstructure*, Wiley.

ii. Heterogeneous Real-Time Trading Strategies in the Foreign Exchange Market; Michel M. Dacorogna, Ulrich A. Müller, Christian Jost, Olivier V. Pictet, Richard B. Olsen and J. Robert Ward, 1995, *The European Journal of Finance*, Vol. 1, p. 383–403.

iii. Correlation of High-Frequency Financial Time Series; Mark Lundin, Michel M. Docoragna, and Ulrich A. Muller; Reprinted from *Financial Markets Tick by Tick.*

iv. Dunis, C. and B. Zhou (eds.), 1998, *Nonlinear Modelling of High Frequency Financial Time Series*, John Wiley & Sons, Chichester.

v. Owain ap Gwilym and Charles Sutcliffe, *High-Frequency Financial Market Data*, p. 63–67, Risk Publications.