High Frequency Quoting:
Short-Term Volatility in Bids and Offers

Joel Hasbrouck

November 13, 2012

I have benefited from the comments received from the Workshop of the Emerging Markets Group (Cass Business School, City University London), SAC Capital, Jump Trading, and Utpal Bhattacharya's doctoral students at the University of Indiana. All errors are my own responsibility. I am grateful to Jim Ramsey for originally introducing me to time-scale decompositions.

Department of Finance, Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012 (Tel: 212-998-0310, jhasbrou@stern.nyu.edu).

High-Frequency Quoting:

Short-Term Volatility in Bids and Offers

Abstract

High-frequency changes, reversals, and oscillations can lead to volatility in a market's bid and offer quotes. This volatility degrades the informational content of the quotes, exacerbates execution price risk for marketable orders, and impairs the reliability of the quotes as reference marks for the pricing of dark trades. This paper examines volatility on time scales as short as one millisecond for the National Best Bid and Offer in the US equity market. On average, in a 2011 sample, volatility at the one millisecond time scale is approximately five times larger than can be attributed to long-term informational volatility. In addition, there are numerous localized episodes involving intense bursts of quote volatility. It is less clear, however, that this volatility can be tied to a recent rise in low-latency technology. Short-term volatility estimated over 2001-2011 historical sample is not characterized by a distinct trend.

I. Introduction

Recent developments in market technology have called attention to the practice of high-frequency trading. The term is used commonly and broadly in reference to all sorts of fast-paced market activity, not just "trades", but trades have certainly received the most attention. There are good reasons for this, as trades signify the actual transfers of income streams and risk. Quotes also play a significant role in trading process, however. This paper accordingly examines short-term volatility in bids and offers of US equities, a consequence of what might be called high frequency quoting.

By way of illustration, Figure 1 depicts the National best bid (NBB) and National best offer (NBO) for AEP Industries (a Nasdaq-listed manufacturer of packaging products) on April 11, 2011. In terms of broad price moves, the day is not a particularly volatile one, and the bid and offer quotes are stable for long intervals. The placidity is broken, though, by several intervals where the bid undergoes extremely rapid changes. The average price levels, before, during and after the episodes are not dramatically different. Moreover, the episodes are largely one-sided: the bid volatility is associated with an only moderately elevated volatility in the offer quote. Nor is the volatility associated with increased executions. These considerations suggest that the volatility is unrelated to fundamental public or private information. It appears to be an artifact of the trading process.

It is not, however, an innocuous artifact. Bids and asks in all markets represent price signals, and, to the extent that they are firm and accessible, immediate trading opportunities. From this perspective, the noise added by quote volatility impairs the informational value of the public price. Most agents furthermore experience latency in ascertaining the location of the bid and offer price and in timing of their order delivery. Elevated short-term volatility increases the execution price risk associated with these delays. In US equity markets the National Best Bid and Offer are particularly important, because they are used as benchmarks to assign prices in so-called dark trades, a category that includes roughly thirty percent of all volume.[1]

---

[1] Dark trading mechanisms do not publish visible bids and offers. They establish buyer-seller matches, either customer-to-customer (as in a crossing network) or dealer-to-customer (as in the case of an internalizing broker-dealer). The matches are priced by reference to the NBBO: generally

In the context of the paper's data sample, the AEPI episode does not represent typical behavior. Nor, however, is it a singular event. It therefore serves to motivate the paper's key questions. What is the extent of short-term volatility? How can we distinguish fundamental (informational) and transient (microstructure) volatility? Finally, given the current public policy debate surrounding low-latency activity, how has it changed over time?

These questions are addressed empirically in a broad sample of US equity market data using summary statistics that are essentially short-term variances of bids and asks. Such constructions, though, inevitably raise the question of what horizon constitutes the "short term" (a millisecond? a minute?). The answer obviously depends on the nature of the trader's market participation, as a collocated algorithm at one extreme, for example, or as a remotely situated human trader at the other. This indeterminacy motivates the use of empirical approaches that accommodate flexible time horizons. One class of standard tools that satisfies this requirement includes methods variously called time-scale, multi-resolution, or wavelet decompositions. The present analysis applies these tools to quote data.[2]

The paper is organized as follows. The next section establishes the economic and institutional motivation for the consideration of local bid and offer variances with sliding time scales. Section III is a short presentation of the essentials of wavelet transformations and time-scale decompositions. The paper then turns to applications. Section IV presents an analysis of a recent sample of US equity data featuring millisecond time stamps. To extend the analysis to historical samples in which time stamps are to the second, Section V describes estimation in a Bayesian framework where millisecond time stamps are simulated. Section VI applies this approach to a historical sample of US data from 2001 to 2011. Connections to high frequency trading and volatility modeling are discussed in Section VII. A summary concludes the paper in Section VIII.

---

at the NBBO midpoint in a crossing network, or at the NBB or the NBO in a dealer-to-customer trade.

[2] Hasbrouck and Saar (2011) examine high-frequency activity within the Inet book. An early draft of the paper used wavelet analyses of message count data to locate periods of intense message traffic.

II. Timing uncertainty and price risk

High frequency quote volatility may be provisionally defined as the short-term variance of the best bid and/or best offer (BBO), that is, the usual variance calculation applied to the BBO over some relatively brief window of time. This section is devoted to establishing the economic relevance of such a variance in a trading context. The case is a simple one, based on: the function and uses of the BBO; the barriers to its instantaneous availability; the role of the time-weighted price mean as a benchmark; and, the interpretation of the variance about this mean as a measure of risk.

In current thinking about markets, most timing imperfections are either first-mover advantages arising from market structure or delays attributed to costly monitoring. The former are exemplified by the dealer's option on incoming orders described in Parlour and Seppi (2003), and currently figure in some characterizations of high-frequency traders (Biais, Foucault and Moinas (2012); Jarrow and Protter (2011)). The latter are noted by Parlour and Seppi (2008) and discussed by Duffie (2010) as an important special case of inattention which, albeit rational and optimal, leads to infrequent trading, limited participation, and transient price effects (also, Pagnotta (2009)).

As a group these models feature a wide range of effects bearing on agents' arrivals and their information asymmetries. An agent's market presence may be driven by monitoring decisions, periodic participation, or random arrival intensity. Asymmetries mostly relate to fundamental (cash-flow) information or lagged information from other markets. Agents in these models generally possess, however, timely and extensive market information. Once she "arrives" in a given market, an agent accurately observes the state of that market, generally including the best bid and offer, depth of the book and so on. Moreover, when she contemplates an action that changes the state of the book (such as submitting, revising or canceling an order), she knows that her action will occur before any others'.

In reality, of course, random latencies in her receipt of information and the transmission of her intentions combine to frustrate these certainties about the market and the effects of her orders. The perspective of this paper is that for some agents these random latencies generate randomness in the execution prices, and that short-term quote variances can meaningfully measure this risk. Furthermore, although all agents incur random latency, the distributions of these delays vary among participants. An agent's latency distribution can be summarized by time-scale, and this in turn motivates time-scale decompositions of bid and offer variances.

While random latencies might well affect strategies of all traders, the present analysis focuses on someone who intends to submit a marketable order (one that seeks immediate execution) or an order to a dark pool. In either case, ignoring hidden orders, an execution will occur at the bid, the offer or at an average of the two. Assume, for the sake of timing notation, that there is one consistent atomic time stamp that is standardized and available throughout the market. Suppose that at time t (a trader transmits a marketable order, but knows that its actual time of arrival at the market is uniformly distributed on $(t, t+\delta]$. The mean and variance of the quote over this interval characterize the first two moments of the distribution of the execution price, conditional on a given price path.

Equivalently, the problem may be viewed as arising from latencies in transmission of the state of the market, where the trader knows her market information represents the state of market at some time in the interval $(t-\delta, t]$. These conjectures are obviously oversimplified. Random transmission latencies undoubtedly exist in both directions, and their distributions are unlikely to be uniform. In these more complicated scenarios, though, the price statistics computed over an interval equal to the average delay should still be useful.

The use of an average price in situations where there is execution timing uncertainty is a common principle in transaction cost analysis. Perold's implementation shortfall measure is usually operationally defined for a buy order as the execution price (or prices) less some hypothetical benchmark price (and for a sell order as the benchmark less the execution price), Perold (1988). As a benchmark price, Perold suggests the bid-ask midpoint prevailing at the time of the decision to trade. Many theoretical analyses of optimal trading strategies use this or similar alternative pre-trade benchmark. Practitioners, however, and many empirical analyses rely on prices averaged over some comparison period. The most common choice is the value-weighted average price (VWAP), although the time-weighted average price (TWAP) is also used. One industry compiler of comparative transaction cost data notes, "In many cases the trade data which is available for analysis does not contain time stamps. …. When time stamps are not available, pension funds and investment managers compare their execution to the volume weighted average price of the stock on the day of the trade" (Elkins-McSherry (2012)). This quote attests to the importance of execution time uncertainty, although a day is certainly too long to capture volatility on the scale of transmission and processing delays. Average prices are also used as objectives by certain execution strategies. A substantial portion of the orders analyzed by Engle, Ferstenberg and Russell (2012) target VWAP, for example.

The situations discussed to this point involve a single trader and single market. In a fragmented market, the number of relevant latencies may be substantially larger. In the US there are presently about 17 "lit" market centers, which publish quotes. A given lit market's quotes are referenced by the other lit markets, dark pools (currently around 30 in number), by executing broker-dealers (approximately 200), and by data consolidators (U.S. Securities and Exchange Commission (2010)). The BBO across these centers, the National Best Bid and Offer (NBBO) is in principle well-defined. The NBBO perceived by any given market center, consolidator or other agent, however, comprises information subject to random transmission delays that differ across markets and receiving agents. These delays introduce noise into the NBBO determination. Local time-averaging (smoothing) can help to mitigate the effects of this noise, while the local variance can help assess the importance of the noise.

As a final consideration, transmission delays can exacerbate the difficulties a customer faces in monitoring the handling of his order. The recent SEC concept release notes that virtually all retail orders are routed to OTC market-makers, who execute the orders by matching the prevailing NBBO (U.S. Securities and Exchange Commission (2010)). Stoll and Schenzler (2006) note that these market-makers may possess a look-back option: To the extent that customers can't verify their order delivery times or the NBBO perceived by the market-maker at the exact time of order arrival, the market-maker possesses flexibility in pricing the execution, and an economic interest in the outcome.

Timing uncertainty may also arise in the mandated consolidated audit trail (U.S. Securities and Exchange Commission (2012)). The rule requires millisecond time-stamps on all events in an order's life cycle (such as receipt, routing, execution and cancellation). This does not suffice to determine the information set (including knowledge of the NBBO) of any particular agent at any particular time. Thus, for example, a dealer's precise beliefs about the NBBO at the time a customer order was received will lie beyond the limits of regulatory verification. It must also be admitted, however, that a system that would permit such a determination in a fragmented market is unlikely to be feasible.

III. Time-scale variance decompositions

This study uses short-term means and variances of bids and offers. Despite the apparent simplicity and directness of such computations, however, it should be noted at the outset that there are two significant departures from usual financial econometrics practices.

Firstly, although prices are assumed to possess a random-walk component (formally, a unit root), the mean and variance calculations are applied to price levels, not first differences. Differencing is sometimes described as high-pass filtering: it maintains the details of the process at the differencing frequency (one millisecond, in this study), but suppresses patterns of longer horizons. For example, if Figure 1 displayed the first difference of the bid instead of the level, the volatility episodes would still be apparent. It would not be obvious, however, that there was no net price change over these episodes. Of course if the price follows a random-walk with drift, a sample mean and variance computed over an interval won't correspond to estimates of a global mean and variance for the process (which doesn't exist). Variances computed over intervals of a given length are nevertheless stationary, however, and amenable to statistical and economic interpretation.

The second point of contrast concerns the interpretation of "short-term". The techniques applied in this study treat the time-scale in a flexible, systematic manner. In most empirical analyses of high-frequency market data, the time scale of the model is determined at an early stage (sometimes by limitations of the data), and the proposed statistical model is parsimonious and of low order with respect to this time scale (for example, a fifth-order vector error correction model applied to bids and asks observed at a one minute frequency). This approach is often perfectly adequate, and it can hardly be considered a devastating criticism to note that such data and models tend to focus on dynamics at a particular time scale and ignore variation over longer and shorter frames. A phenomenon like high-frequency quoting, however, does not present an obvious choice for time scale. It is therefore advantageous to avoid that choice, and pursue an empirical strategy that treats all time scales in a unified manner.

To this end, wavelet transformations, also known as time-scale or multi-resolution decompositions are widely used across many fields. The summary presentation that follows attempts to cover the material only to a depth sufficient to understand the statistical evidence marshaled in this study. Percival and Walden (2000, henceforth PW) is comprehensive textbook presentation that highlights the connections to conventional stationary time series analysis. The present notation closely follows PW. Gençay, Selçuk and Whitcher (2002) discuss economic and financial applications in the broader context of filtering. Nason (2008) discusses time series and other applications of wavelets in statistics. Ramsey (1999) and Ramsey (2002) provides other useful economic and financial perspectives. Walker (2008) is clear and concise, but oriented more toward engineering applications.

Studies that apply wavelet transforms to the economic analysis of stock prices loosely fall into two groups. The first set explores time scale aspects of stock comovements. A stock's beta is a summary statistic that reflects short-term linkages (like index membership or trading-clientele effects) and long-term linkages (like earnings or national prosperity). Wavelet analyses can characterize the strength and direction of these horizon-related effects (for example, Gençay, Selçuk and Whitcher (2002); In and Kim (2006)). Most of these studies use wavelet transforms of stock prices at daily or longer horizons. A second group of studies uses wavelet methods to characterize volatility persistence (Dacorogna, Gencay, Muller, Olsen and Pictet (2001); Elder and Jin (2007); Gençay, Selçuk, Gradojevic and Whitcher (2009); Gençay, Selçuk and Whitcher (2002); Høg and Lunde (2003); Teyssière and Abry (2007)). These studies generally involve absolute or squared returns at minute or longer horizons. Wavelet methods have also proven useful for jump detection and jump volatility modeling Fan and Wang (2007). Beyond studies where the focus is primarily economic or econometric lie many more analyses where wavelet transforms are employed for ad hoc stock price forecasting (Atsalakis and Valavanis (2009); Hsieh, Hsiao and Yeh (2011), for example).

III.A. The intuition of wavelet transforms: a microstructure perspective

A wavelet transform represents a time series in terms of averages and differences in averages constructed over intervals of systematically varying lengths. By way of illustration, consider a non-stochastic sequence of eight consecutive prices: $p = [p_0 \quad p_1 \quad \cdots \quad p_7]$. A trader whose order arrival time is random and uniformly distributed on this set expects to trade at the overall mean price, $\mu = \frac{1}{8}\sum_t p_t$. In the terminology of wavelet transforms, a mean computed over an interval is a *wavelet smooth*. The *time scale* of the smooth is the horizon over which it is considered constant (eight, in this case). The *level* of the smooth is $j = \log_2(time\ scale)$; $j = 3$ in this case $(2^3 = 8)$. The choice of sample length as an integer power of two is deliberate; generalizations will shortly be indicated. The top-level smooth is $\mathcal{S}_{j=3} = [\mu \quad \mu \quad \cdots \quad \mu]$, that is, a row vector consisting of the mean repeated eight times (to conform to the original price vector). The deviations from the mean define the *wavelet rough,* $\mathcal{R}_3 = p - \mathcal{S}_3$. The variance $Var(\mathcal{R}_3)$ indicates the risk or uncertainty faced by this trader.

The sequence might also be considered from the perspective of a faster trader who might be randomly assigned to trade in the first or the last half of the sequence, in $\{0, 1, 2, 3\}$ or $\{4, 5, 6, 7\}$, but within each of these sets faces order arrival uncertainty of length four. Her benchmark prices are defined by the smooth

$$\mathcal{S}_2 = \left[\frac{p_0 + \cdots + p_3}{4} \quad \frac{p_4 + \cdots + p_7}{4}\right] \otimes [1 \quad 1 \quad 1 \quad 1] \tag{1}$$

where $\otimes$ denotes the Kronecker product: $\mathcal{S}_2$ is the mean of the first four values repeated four times joined to the mean of the second four values repeated four times. Each of the means is constant over a time scale of $2^2 = 4$. The corresponding rough at this level is $\mathcal{R}_2 = p - \mathcal{S}_2$. Finally consider a still-faster trader who is randomly assigned to one of the four intervals $\{0,1\}, \{2,3\}, \{4,5\}$ or $\{6,7\}$, and within each interval faces random arrival over an interval of length two. The corresponding smooth is

$$\mathcal{S}_1 = \left[\frac{p_0 + p_1}{2} \quad \frac{p_2 + p_3}{2} \quad \frac{p_4 + p_5}{2} \quad \frac{p_6 + p_7}{2}\right] \otimes [1 \quad 1] \tag{2}$$

Each of the four two-period means is constant over a time scale of $2^1 = 2$. The rough is $\mathcal{R}^1 = p - \mathcal{S}_1$. In all we have three decompositions embodying different time scales. The rough variances indicate the uncertainties faced by traders at each time scale.

There is another way of looking at these decompositions. If the top-level smooth $\mathcal{S}_3$ captures all variation at time scales of eight (and higher, if we allow the sequence to be embedded in a larger sample), then the corresponding rough $\mathcal{R}_3$ must capture variation at time scales four and lower. Similarly, the rough $\mathcal{R}_2$ must capture variation at time scales of two and lower. The difference between them defines the *detail* component $\mathcal{D}_3 = \mathcal{R}_3 - \mathcal{R}_2$, which captures variation on a scale of four (only). Similarly, $\mathcal{D}_2 = \mathcal{R}_2 - \mathcal{R}_1$ captures variation on a time scale of two; $\mathcal{D}_1 = \mathcal{R}_1$ captures variation on a time scale of one. The time scale of detail component $\mathcal{D}_j$ is denoted $\tau_j = 2^{j-1}$.

Thus, in addition to the rough/smooth decompositions, we have a series of detail/smooth decompositions: $p = \mathcal{D}_1 + \mathcal{D}_2 + \mathcal{D}_3 + \mathcal{S}_3 = \mathcal{D}_1 + \mathcal{D}_2 + \mathcal{S}_2 = \mathcal{D}_1 + \mathcal{S}_1$. The advantage of the rough/smooth decompositions is that they correspond more closely to components of economic interest (the risk faced by traders at a particular and shorter time scales). The advantage of the detail/smooth decompositions is that they can be shown to be orthogonal: $\mathcal{D}_i \cdot \mathcal{D}_j = 0$ for $i \neq j$. This orthogonality facilitates clean time-scale decompositions of variances.

The progression from coarser to finer time scales in this illustration follows the approach of an econometrician who summarizes the coarser features of a data set before moving on to the finer features. Most wavelet computations, though, including the standard pyramid algorithm, are implemented in the opposite direction, from fine to coarse.

The averages used in this example are simple arithmetic means. The process of generating these means at various time scales is formally called a discrete Haar transform. Alternative discrete wavelet transforms (DWTs) are generated by weighting the means in various ways. The discrete Haar transform is easy to generalize to any sequence of dyadic (integer power of two) length, but few data samples are likely to satisfy this requirement. A further drawback is that the transform is also sensitive to alignment. For example, if we rotate the price sequence one position, obtaining $[p_7 \quad p_0 \quad \cdots \quad p_6]$, the details, smooths and roughs are not correspondingly rotated.

The maximal overlap discrete wavelet transform (MODWT) is an alternative transform that fixes the alignment sensitivity and the power-of-two sample size limitation (PW, Ch. 5). In the MODWT, the detail and smooth components are averaged over all cyclic permutations. This is an accepted and widely-used approach, but it comes at the cost of orthogonality. Notationally indicating the MODWT by a tilde "∼", $\widetilde{D}_1$ and $\widetilde{D}_2$ are not orthogonal, and $p \cdot p \neq \widetilde{\mathcal{D}}_1 \cdot \widetilde{\mathcal{D}}_1 + \widetilde{\mathcal{D}}_2 \cdot \widetilde{\mathcal{D}}_2 + \tilde{\mathcal{S}}_1 \cdot \tilde{\mathcal{S}}_1$. Sum-of-squares decompositions are still achievable under the MODWT (we can still compute the variances of smooths, roughs, and details), but these must be computed from the wavelet transform coefficients.

III.B. Time-scale decompositions of difference-stationary processes

In the example of the last section the price sequence is non-stochastic, and all of the randomness resides in the order arrival time. This device allows us to compute and interpret the means and variances implied by the wavelet transform without reference to the price dynamics. In this section we allow the price to follow a stochastic process. Order arrival randomness still serves to motivate interest in the wavelet means and variances, but this arrival process is not explicitly discussed.

The price process is assumed to be first-difference-stationary, which accommodates the usual basic framework of an integrated price with stationary first differences. Note that despite the presence of the random-walk component, we compute the transforms of price levels, rather than first differences.

The wavelet variance at time scale τ_j is denoted $\nu^2(\tau_j)$. For the DWT described here, $\nu^2(\tau_j) = Var(\mathcal{D}_j)$, and the orthogonal sum-of-squares decomposition implies a parallel decomposition of sample variance. As in the discussion above, the variances of the wavelet roughs figure prominently in characterizing time-related execution price risk. They can be computed as $Var(\mathcal{R}_j) = \sum_{i=1}^{j} \nu^2(\tau_j)$. For the MODWT, the wavelet variances can't be computed directly from the detail and smooth components, but they can be computed from the wavelet coefficients (PW Ch. 8). Wavelet variance, covariance and correlation estimates based on MODWT's of bid and ask quotes are the foundation of the analysis.

III.C. Variance ratios

A long tradition in empirical market microstructure assesses the strength of microstructure effects using ratios that compare a short-term return variance to a long-term return variance

(Amihud and Mendelson (1987); Barnea (1974); Hasbrouck and Schwartz (1988)).[3] The idea is that microstructure imperfections introduce mispricing that inflates short-term variance relative to long-term fundamental variance. Ratios constructed from wavelet variances give a more precise and nuanced characterization of this effect because the long-term wavelet variance is effectively stripped of all short-term components, and the short-term wavelet variance can focus on a particular time scale.

Suppose that a price evolves according to: $p_t = p_{t-1} + \epsilon_t$, where $E\epsilon_t = 0, E\epsilon_t^2 = \sigma_\epsilon^2$ and $E\epsilon_t\epsilon_s = 0$ for $t \neq s$. A conventional variance ratio for horizons $m, n > 0$ might be defined as:

$$\frac{Var(p_t - p_{t-n})/n}{Var(p_t - p_{t-m})/m} = \frac{Var(p_t - p_{t-n})}{Var(p_t - p_{t-m})} \times \frac{m}{n} = 1 \tag{3}$$

If $n < m$, ratio estimates in excess of one indicate inflated short-term volatility. The n/m term essentially normalizes the variances to the benchmark random walk.

A wavelet variance ratio is defined in a similar fashion, but with a different normalization term. PW (p. 337) show that for this random-walk the (Haar) wavelet variances of p are:

$$\nu_{rw}^2(\tau_j) = \frac{\sigma_\epsilon^2}{6}\left(\tau_j + \frac{1}{2\tau_j}\right) \tag{4}$$

With this result it is natural to define a wavelet variance ratio as:

$$VR_{n,m} = \frac{\nu^2(\tau_n)}{\nu^2(\tau_m)} \times \frac{\nu_{rw}^2(\tau_m)}{\nu_{rw}^2(\tau_n)} \tag{5}$$

The ν_{rw}^2 divisors normalize the price wavelet variances similar to role of m/n in the conventional variance ration; the σ_ϵ^2 parameter cancels. If the price process is a random walk, the wavelet variance ratio is unity. More generally, deviation from unity measures excess or subnormal volatility.

---

[3] Return variance ratios are also used more broadly in economics and finance to characterize deviations from random-walk behavior over longer horizons (Charles and Darné (2009); Faust (1992); Lo and MacKinlay (1989)).

III.D. Extensions to coarser time scales

The wavelet transforms in the present analysis are performed at a one-millisecond resolution. This is necessary to capture the high-frequency phenomena of primary interest. It is also useful, however, to measure relatively longer components, on the order of thirty minutes or so. These components can be computed directly from the one-millisecond data, but the computations are lengthy and burdensome. Instead the longer-horizon calculations are performed with a one-second resolution. For these calculations, the millisecond prices are first averaged over each second, and the wavelet transforms are computed for the resulting series of one-second average prices. The corresponding wavelet variances at level *j* for these average prices are denoted $\nu_a^2(\tau_{a,j})$ where the time scale in milliseconds is $\tau_{a,j} = 1{,}000 \times 2^{j-1}$.

Under the assumption that the price follows a one-millisecond random-walk (that is, $p_t = p_{t-1} + \epsilon_t$ where *t* indexes milliseconds), the sequence of one-second average prices is integrated with autocorrelated differences, an IMA(1,1) process. The wavelet variances are of the form $\nu_{a,rw}^2(\tau_{a,j}) = \alpha_j \sigma_\epsilon^2$, where (as above) a indicates the pre-transform-averaging and α_j are proportionality factors. The α_j not have a simple closed-form representation (as in equation (4)), but they can be computer numerically. With these results, it is natural to construct a variance ratio that uses the finer (one millisecond) resolution for the smaller time scales and the coarser (one second) resolution for the longer time scales:

$$VR_{n,m} = \frac{\nu^2(\tau_n)}{\nu_a^2(\tau_{a,m})} \times \frac{\nu_{a,rw}^2(\tau_{a,m})}{\nu_{rw}^2(\tau_n)} \tag{6}$$

III.E. Estimation

Estimates of wavelet variances and related quantities are basically formed as sample analogues of population parameters. PW discuss computation and asymptotic distributions. In most applications, the wavelet variance estimate at a particular time scale is computed from the transformation of the full data series. In the present case, these estimates are formed over fifteen-minute subintervals. There are several reasons for this. Firstly, it yields computational simplifications. Secondly, subinterval calculations can help characterize the distribution of the variance estimates. (PW suggest this for large samples, p. 315.) Thirdly and most importantly, though, it offers a quick and approximate way to accommodate nonstationarity. The paper's opening example suggests that high-frequency quoting might involve localized bursts. Interval-based variance measures offer a simple way to detect these bursts.[4] Figure 1 (and similar episodes) were located in this manner.

IV. A cross-sectional analysis

From a trading perspective, stocks differ most significantly in their general level of activity (volume measured by number of trades, shares or values). The first analysis aims to measure the general level of HFQ volatility and to relate the measures to trading activity in the cross-section for a recent sample of firms.

IV.A. Data and computational conventions.

The analyses are performed for a subsample of US firms using trading data from April, 2011 (the first month of my institution's subscription.) The subsample is constructed from all firms present on the CRSP and TAQ databases from January through April of 2011 with share codes of 10, 11, or 12, and with a primary listing on the New York, American or Nasdaq exchanges. I compute the daily average dollar volume based on trading in January through March, and form decile rankings on this variable. Within each decile I sort by ticker symbol and take the first ten firms. Table 1 reports summary statistics, with subsamples are grouped into quintiles for brevity.

The quote data are from the NYSE's Daily TAQ file and constitute the consolidated quote feed for all stocks listed on a US exchange, with millisecond time-stamps.[5] A record in the consolidated quote (CQ) file contains the latest bid and offer originating at a particular exchange. If the bid and offer establish the best in the market (the "National Best Bid and Offer," NBBO) this fact is noted on the record. If the CQ record causes the NBBO to change for some other reason, a

---

[4] Wavelet transformations are widely used in noise detection and signal de-noising. These techniques are certainly promising tools in the study of microstructure data. For present purposes, though, the simpler approach of interval computations suffices.

[5] The "daily" reference in the Daily TAQ dataset refers to the release frequency. Each morning the NYSE posts files that cover the previous day's trading. The Monthly TAQ dataset, more commonly used by academics is released with a monthly frequency and contains time stamps in seconds.

message is posted to another file (the NBBO file). Thus, the NBBO can be obtained by merging the CQ and NBBO files. It can also be constructed (with a somewhat more involved computation) directly from the CQ file. Spot checks verified that these two approaches were consistent.

The NBB and NBO are usually valid continuously from approximately 9:30 to 16:00 ("normal" US trading hours). It is well known, however, that volatility is elevated at the start and finish of these sessions. This is particularly acute for low-activity firms. In these issues, sessions may start with wide spreads, which subsequently narrow appreciably before any trades actually occur. To keep the analysis free of this starting and ending volatility, I restrict the computations to the interval 9:45 to 15:45.

For time scales ranging from one millisecond to 32.8 seconds ($\approx 2^{15}$ milliseconds), wavelet transformations are computed on a one millisecond grid. With this resolution each day's analysis interval contains $6 \times 3{,}600 \times 1{,}000 \approx 2 \times 10^7$ observations (for each stock). For computational expediency, transformations on a one-second grid are computed for time scales of two seconds to 2^{11} seconds $= 34.1$ minutes. The overlap in time scales for the millisecond and second analyses serves as a computational check. To facilitate comparison of these analyses, the one-second prices are computed as averages of the one-millisecond prices (as opposed to, say, the price prevailing at the end of the second).

IV.B. Rough variances

As discussed in Sections II and III.A, the rough variance $Var(\mathcal{R}_j)$ measures the execution price uncertainty faced by trader with arrival time uncertainty at time scales τ_j and shorter. The wavelet transforms are computed for bids and offers stated in dollars per share. This is meaningful because many trading fees (such as commissions and clearing fees) are assessed on a per share basis. Access fees, the charges levied by exchanges on taker (aggressor) sides of executions are also assessed per share. US SEC Regulation NMS caps access fees at 3 mils ($0.003) per share, and in practice most exchanges are close to this level. Practitioners regard access fees as significant to the determination of order routing decisions, and this magnitude therefore serves as a rough threshold of economic importance.

Table 2 (Panel A) presents estimates for $\sqrt{Var(\mathcal{R}_j)}$ (that is, the standard deviation) in units of mils per share. For brevity, the table does not report estimates for all time scales. It will be recalled that the rough variance at a given time scale also impounds variation due to components at all shorter time scales. For example, the standard deviation at the 64 millisecond time scale also captures variation at 32, 18, 8, 4, 2 and 1 millisecond time scales. Relative to access fees (three mils per share), short-term volatility is not particularly high. The access fee threshold is not reached in the full sample until the time-scale is extended to 4.1 seconds. At the lowest reported time scale (64 milliseconds and below) the average volatility is only 0.4 mils.

Most analyses involving investment returns or comparison across firms assume that share normalizations are arbitrary. From this perspective, it is sensible to normalize the rough variances by price per share. Table 2 Panel B reports estimates of $\sqrt{Var(\mathcal{R}_j)}/\bar{p}$, where $\bar{p}$ is the average bid-offer midpoint over the estimation interval, in basis points (0.01%). By this measure, too, volatility at the shortest time scale appears modest, 0.3 bp on average.

In comparing the two sets of results, it appears that the basis point volatilities decline by a factor of roughly five in moving from the lowest to the highest dollar volume quintiles (Table 2, Panel B). Most of this decline, though, is due to the price normalization. From Table 1, the price increases by a factor of about ten over the quintiles. The volatilities in mils per share (Table 2, panel A) increase, but only by a factor of around two. This relative constancy suggests that quote volatility is best characterized as a "per share" effect, perhaps due to the constancy of the tick size or the prevalence of per-share cost schedules.

IV.C. Wavelet variance ratios

Wavelet variance ratios normalize short-term variances by long-term variances under a random-walk assumption. Table 3 reports within-quintile means and standard errors. (The standard error computations assume independence across observations.) Figure 2 presents the means graphically, as a function of time scale. The variance ratios are normalized with respect to variation at the longest time scale in the analysis (2^{11} seconds = 34.1 minutes). The ratio at 34.1 minutes is therefore unity by construction. If price dynamics followed a random walk, the variance ratios would be unity at all time scales, and the plots in the figure would be flat.

The table and figure summarize the results of analyses at a one-millisecond resolution, and (for the longer time scales) analyses of one-second averaged prices. From time scales of roughly four to sixty seconds, these two computations overlap, and at time scales that are approximately equal the two computations are in close agreement.

The overall sample averages (first column) suggest substantial excess short-term volatility. The value of 5.36 at a one millisecond time-scale simply implies that volatility at this time-scale is over five times higher than would be implied by a random walk calibrated to 34.1 minute volatility. In the lowest two dollar volume quintiles, the volatility is inflated by a factor of approximately nine. The estimate for the highest dollar volume quintile is somewhat lower (at 1.83), but still implies a volatility inflation of 80 percent.

IV.D. Bid and offer correlations by time scale

The excess short-term volatility indicated by the high variance ratios (Table 3) suggests that the volatility is not of a fundamental or informational nature. Additional evidence on this point is suggested by examination of the correlations between bid and offer components on various time scales. Table 4 presents these estimates with standard errors; Figure 3 depicts the correlations graphically. For brevity, Table 4 does not present estimates based on the one-second resolution analysis over the time-scales where they overlap with the millisecond resolution analysis. The figure depicts all values, which visually confirms the consistency of the two sets of estimates in the overlap region.

Hansen and Lunde note that to the extent that volatility is fundamental, we would expect bid and offer variation to be perfectly correlated, that is, that a public information revelation would shift both prices by the same amount Hansen and Lunde (2006). Against this presumption, the short-term correlation estimates are striking. At time scales of 128 ms or lower, the correlation is below 0.7 for all activity quintiles. For the shortest time scales and lower activity quintiles, the correlation is only slightly positive. This suggests that substantial high-frequency quote volatility is of a dinstinctly transient nature.

V. Time-scale decompositions with truncated time stamps.

The analysis in the preceding section relies on a recent one-month sample of daily TAQ data. For addressing policy issues related to low-latency activity, it would be useful to conduct a historical analysis, spanning the period over which low-latency technology was deployed. Extending the analysis backwards, however, is not straightforward. Millisecond time-stamps are only available in the daily TAQ data from 2006 onwards. Monthly TAQ data (the standard source used in academic research) is available back to 1993 (and the precursor ISSM data go back to the mid-1980s). These data are substantially less expensive than the daily TAQ, and they have a simpler logical structure.

The time stamps on the Monthly TAQ and ISSM datasets are reported only to the second. At first glance this might seem to render these data useless for characterizing sub-second variation. This is unduly pessimistic. It is the purpose of this section to propose, implement and validate an approach for estimating sub-second (indeed, millisecond) characteristics of the bid and ask series using the second-stamped data. This is possible because the data generation and reporting process is richer than it initially seems.

Specifically, the usual sampling situation in discrete time series analysis involves either aggregation over periodic intervals (such as quarterly GDP) or point-in-time periodic sampling (such as the end-of-day S&P index). In both cases there is one observation per interval, and in neither case do the data support resolution of components shorter than one interval. In the present situation, however, quote updates occur in continuous time and are disseminated continuously. The one second time-stamps arise as a truncation (or equivalently, a rounding) of the continuous event times. The Monthly TAQ data include all quote records, and it is not uncommon for a second to contain ten or even a hundred quote records.

Assume that quote updates arrive in accordance with a Poisson process of constant intensity. If the interval $(0, t)$ contains n updates, then the update times have the same distribution as the order statistics corresponding to n independent random variables uniformly distributed on the interval $(0, t)$ (Ross (1996), Theorem 2.3.1). Within a one-second interval containing n updates, therefore, we can simulate continuous arrival times by drawing n realizations from the standard uniform distribution, sorting, and assigning them to quotes (in order) as the fractional portions of the arrival times. These simulated time-stamps are essentially random draws from true distribution. This result does not require knowledge of the underlying Poisson arrival intensity.

We make the additional assumption that the quote update times are independent of the updated bid and ask prices. (That is, the "marks" associated with the arrival times are independent of the times.) Then the wavelet transformations and computations on the time-stamp-simulated series constitute a draw from their corresponding posterior distributions.

This estimation procedure can be formalized in a Bayesian Markov-Chain Monte Carlo (MCMC) framework. To refine the estimates, we would normally make repeated iterations ("sweeps") over the sample, simulating the update times, computing the wavelet transforms, and It also bears mention that bid and ask quotes are paired. That is, a quote update with a time-stamp of 9:30:01 contains both a bid and ask price. We may not know exactly when within the second the update occurred, but we do know that the bid and ask were updated (or refreshed, if not changed) at the same time. This alignment strengthens the inferences about the wavelet correlations.

It is readily granted that few of the assumptions underlying this model are completely satisfied in practice. For a time-homogeneous Poisson process, inter-event durations are independent. In fact, inter-event times in market data frequently exhibit pronounced serial dependence, and this feature is a staple of the autoregressive conditional duration and stochastic duration literature (Engle and Russell (1998); Hautsch (2004)). In Nasdaq Inet data, Hasbrouck and Saar (2011) show that event times exhibit intra-second deterministic patterns. Suboordinated stochastic process models of security prices suggest that transactions (not wall-clock time) are effectively the "clock" of the process (Shephard (2005)).

There exists, however, a simple test of the practical adequacy of the randomization procedure. The time-stamps of the data analyzed in the last section are stripped of their millisecond remainders. New millisecond remainders are simulated, the random-time-stamped data are analyzed, and we examine the correlations between the two sets (original and randomized) of estimates. Table 5 summarizes the cross-firm distribution of these correlations. For the wavelet variances, the agreement between original and randomized estimates is very high for all time scales and in all subsamples. Even at the briefest time scale of one millisecond, the median correlation is 0.991. At time-scales of one second and above, the agreement is near perfect.

Given the questionable validity of some of the assumptions, and the fact that only one draw is made for each second's activity, this agreement might seem surprising. It becomes more reasonable, however, when one considers the extent of averaging underlying the construction of both original and randomized estimates. There is explicit averaging in that each wavelet variance estimate formed over a fifteen-minute interval involves (with a millisecond resolution) 900,000 inputs. As long as the order is maintained, a small shift in a data point has little impact over the overall estimate. Finally, inherent in the wavelet transformation is an (undesirable) averaging across time scales known as leakage (PW, p. 303).

Agreement between original and randomized bid-ask correlations is weaker, although still under the circumstances, quite good. The median correlation of one millisecond components is 0.219 (in the full sample), but this climbs to 0.577 at a time scale of 128 ms. The reason for the poorer performance of the randomized correlation estimates is simply that the wavelet covariance between two series is sensitive to relative alignment. When a bid change is shifted even by a small amount relative to the offer, the inferred pattern of comovement is distorted.

Across dollar volume quintiles, the correlations generally improve for all time scales. This is true for both wavelet variances and correlations, but is more evident in the latter. This is a likely consequence of the greater incidence, in the higher quintiles, of multiple quote records within the same second. Specifically, for a set of *n* draws from the uniform distribution, the distribution of any order statistic tightens as *n* increases. (For example, the distribution of the 499^{th} order statistic in a sample of 500 in a given second is tighter than the distribution of the first order statistic in a sample of one.) Essentially, an event time can be located more precisely within the second if the second contains more events. This observation will have bearing on the analysis of historical samples with varying numbers of events.

In working with Monthly TAQ data, Holden and Jacobsen (2012, HJ) suggest assigning sub-second time stamps by evenly-spaced interpolation. If there is one quote record in the second, it is assigned a millisecond remainder of 0.500 seconds; if two records, 0.333 and 0.667 seconds, and so on. HJ show that interpolation yields good estimates of effective spreads. It is not, however, equivalent to the present approach. Consider a sample in which each one-second interval contains one quote record. Even spacing places each quote at its half-second point. As a result, the separation between each quote is one second. For example, a sequence of second time stamps such as 10:00:01, 10:00:02, 10:00:03 … maps to 10:00:01.500, 10:00:02.500, 10:00:03.500, and so on. The interpolated time stamps are still separated by one second, and therefore the sample has no information regarding sub-second components. In contrast, a randomized procedure would sweep the space of all possibilities, including 10:00:01.999, 10:00:02.000, …, which provides for attribution of one-millisecond components. Of course, as the number of events in a given one-second interval increases, the two approaches converge: the distribution of the *kth* order statistic in a sample of n uniform observations collapses around its expectation, $k/(n+1)$ as n increases.

For one class of time-weighted statistics in this setting, interpolated time stamps lead to unbiased estimates. Consider a unit interval where the initial price, p_0, is known, and there are n subsequent price updates $p_i, i = 1, \dots, n$ at occurring at times $0 < t_1 < \cdots < t_n < 1$. The time-weighted average of any price function $f(p)$ is $Avg^{TW} = \sum_{i=0}^{n} f(p_i)(t_{i+1} - t_i)$, where $t_0 \equiv 0$ and $t_{n+1} \equiv 1$. Assuming a time-homogeneous Poisson arrival process, the t_i are distributed (as above) as uniform order statistics. This implies $Et_i = i/(n+1)$, the linear interpolated values. If the marks (the p_i) are distributed independently of the t_i, $E[Avg^{TW}] = (n+1)^{-1} \sum_{i=0}^{n} f(p_i)$. This result applies to time-weighted means of prices and spreads (assuming simultaneous updates of bids and offers). It also applies to wavelet transforms and other linear convolutions. It does not apply to variances (or wavelet variances), however, which are nonlinear functions of arrival times.

VI. Historical evidence

This section describes the construction and analysis of variance estimates for a sample of US stocks from 2001 to 2011. In each year, I construct variance estimates for a single representative month (April) for a subsample of firms.

The historical span is problematic in some respects. The period covers significant changes in market structure and technology. Decimalization had been mandated, but was not completely implemented by April, 2001. Reg NMS was adopted in 2005, but was implemented in stages. Dark trading grew over the period. Market information and access systems were improved, and latency emerged as a key concern of participants. The period also includes many events related to the financial crisis, which are relatively exogenous to equity market structure.

The net effect of these developments as they pertain to the present study is that it can safely be asserted that over the period the very nature of bid and ask quotations changed. Markets in 2001 were still dominated by what would later be called "slow" procedures. Quotes were often set manually. Opportunities for automated execution against these quotes were limited (cf. the NYSE's odd-lot system, and Nasdaq's Small Order Execution System). With the advent of Reg NMS, the NBBO became much more accessible (for automated execution).

VI.A. Data

The data for this phase of the analysis are drawn from CRSP and *Monthly* TAQ datasets. In each year, from all firms present on CRSP and TAQ in April, with share codes in (10, 11, 12), and with primary listings on the NYSE, American and Nasdaq exchanges, I draw a subsample of thirty firms. The sampling scheme is random, and stratified by market capitalization.[6] Quote data are drawn from TAQ.

Table 6 reports summary statistics. The oft-remarked increase in the intensity of trading activity is clearly visible in the trends for median number of trade and quote records. From 2001 to 2011, the average compound growth rate in trades is about 26 percent. The average compound growth rate in quotes is about 31 percent.

As described in the last section, all of a firm's quote records in a given second are assigned random, but order preserving, millisecond remainders. The NBBO is constructed from these quote records. This yields a NBBO series with (simulated) millisecond time stamps. From this point, calculation of wavelet transformations and normalizations follows the procedure described in the cross-sectional analysis.

VI.B. Results

The presentation of results largely parallels that of the cross-sectional analysis, except that the variation is across time instead of the cross-section. Panel A of Table 7 summarizes select rough volatilities in mils per share. There is certainly variation from year-to-year, but no time scale

---

[6] As of April, 2001, Nasdaq had not fully implemented decimalization. For this year, I do not sample from stocks that traded in sixteenths.

suggests an increasing trend. Panel B of Table 7 summarizes the price-normalized volatilities in basis points. There is more variation here, and one year in the latter part of the sample (2009) has the highest value. This year also has the lowest average share price, however, suggesting that the 2009 value is mostly an artifact of the normalization.

Table 8 summarizes the wavelet variance ratios. Under the null hypothesis of a homoscedastic random walk, these should be unity for all time scales. As in the cross-sectional analysis, however, they are substantially in excess of one at the shorter time-scales. The general behavior across time is suggestive, but not definitive. For sub-second time scales, the first two years (2001 and 2002) generally have the lowest ratios. All ratios are generally largest for 2010, but are lower in 2011. (Recall that by construction, these are normalized to 34-minute volatility in each year, a procedure that should in principle control for variation in fundamental volatility.) Thus the overall picture depicts a roughly increasing volatility, but the trend is certainly not monotonic and standard errors are large.

Given the media attention devoted to low-latency activity and the undeniable growth in quote volume, the absence of a strong trend in quote volatility seems surprising. There are several possible explanations. In the first place, "flickering quotes" drew comment well before Reg NMS and during time when quotes were dominated by human market makers (Harris (1999); U.S. Commodities Futures Trading Commission Technology Advisor Committee (2001)). The practice of "gapping" the quotes is also an artifact of this era (Jennings and Thirumalai (2007)). In short, the quotes may have in reality been less unwavering than popular memory holds. The apparent discrepancy between quote volatility and quote volume can be explained by appealing to the increase in market fragmentation and consequent growth in matching quotes.

The introductory example of AEPI quotes noted the abrupt starting and stopping of extreme quote volatility periods. A final possibility then is that while quote volatility has not increased on average, there is an increased incidence of extreme, but brief, episodes. Finding (or not finding) these episodes seems to be a fruitful area for further research, and the localized nature of wavelet transforms suggests that they will be useful tools in this search.

VII. Discussion

From an economic perspective, high frequency quote volatility is connected most closely to other high frequency and low latency phenomena in modern markets. From a statistical perspective, it is connected to volatility modeling.

VII.A. High-frequency quoting and high-frequency trading

Most definitions of algorithmic and high-frequency trading encompass many aspects of market behavior (not just executions), and would be presumed to cover quoting as well.[7] Executions and quotations are nevertheless very different events. It is therefore useful to consider their relation in the high-frequency context.

The discussion in section II associates short-term quote volatility with price uncertainty for those who submit marketable orders, use dark mechanisms that price by reference, or simply monitor their brokers. From this perspective, quote volatility is an inverse measure of market quality. It is not necessarily associated with high-frequency executions. One can envision regimes where relatively stable quotes are hit with extreme alacrity when fundamental valuations change, and periods (such as Figure 1) where frenetic quoting occurs in the absence of executions. Nevertheless, the same technology that makes high-frequency executions possible also facilitates the rapid submission, cancellation and repricing of the nonmarketable orders that define the bid and offer. One might expect this commonality of technology to link the two activities in practice.

Executions are generally emphasized over quotes when identifying agents as high-frequency traders. For example, Kirilenko, Kyle, Samadi and Tuzun (2010) select on high volume and low inventory. The low inventory criterion excludes institutional investors who might use algorithmic techniques to accumulate or liquidate a large position. The Nasdaq HFT dataset uses similar criteria (Brogaard (2010); Brogaard, Hendershott and Riordan (2012)). Once high-

---

[7] A CFTC draft definition reads: "High frequency trading is a form of automated trading that employs: (a) algorithms for decision making, order initiation, generation, routing, or execution, for each individual transaction without human direction; (b) low-latency technology that is designed to minimize response times, including proximity and co-location services; (c) high speed connections to markets for order entry; and (d) high message rates (orders, quotes or cancellations)" (U.S. Commodities Futures Trading Commission (2011)).

frequency traders are identified, their executions and the attributes of these executions lead to direct measures of HF activity in panel samples.

In some situations, however, identifications based on additional, non-trade information are possible. Menkveld (2012) identifies one Chi-X participant on the basis of size and prominence. The Automated Trading Program on the German XETRA system allows and provides incentives for designating an order as algorithmic (Hendershott and Riordan (2012)). Other studies analyze indirect measures of low-latency activity. Hendershott, Jones and Menkveld (2011) use NYSE message traffic. Hasbrouck and Saar (2011) suggest strategic runs (order chains) of cancel and replace messages linked at intervals of 100 ms or lower.

Most of these studies find a positive association between low-latency activity and common market quality measures, such as posted and effective spreads. Most also find a zero or negative association between low-latency activity and volatility, although the constructed volatility measures usually span intervals that are long relative to those of the present paper. With respect to algorithmic or high-frequency activity: Hendershott and Riordan (2012) find an insignificantly negative association with the absolute value of the prior 15-minute return; Hasbrouck and Saar (2011) find a negative association with the high-low difference of the quote midpoint over a 15-minute interval; Brogaard (2012) finds a negative relation with absolute price changes over intervals as short as ten seconds.

The time-scaled variance estimates used here clearly aim at a richer characterization of volatility than the high/low or absolute return proxies used in the studies above. The present study does not, on the other hand, attempt to correlate the variance measures with intraday proxies for high-frequency trading. Nevertheless, viewed broadly, the conclusions are consistent in that the period of time spanning the rise of high-frequency trading is not associated with an increasing trend in short term volatility.

One would naturally assume, of course, that the ultimate strategic purpose of high-frequency quoting is to facilitate a trade or to affect the price of a trade. The mechanics of this are certainly deserving of further research.

VII.B. High-frequency quoting and volatility modeling

Security prices at all horizons are a mix of integrated and stationary components. The former are usually identified with persistent fundamental information innovations; the latter, with transient microstructure effects. The former are important to long-term hedging and investment; the latter, to trading and market-making. The dichotomy is sometimes reflected in different statistical tools and models.

Between the two approaches, the greatest common concerns arise in the analysis of realized volatility (Andersen, Bollerslev, Diebold and Ebens (2001); Andersen, Bollerslev, Diebold and Labys (2003a); Andersen, Bollerslev, Diebold and Labys (2003b)). RVs are calculated from short-term price changes. They are useful as estimates of fundamental integrated volatility (IV), and typically serve as inputs to longer-term forecasting models. RVs constructed directly from trade, bid and offer prices are typically noisy, however, due to the presence of microstructure components. Local averaging moderates these effects. The issues are surveyed in Hansen and Lunde (2006) and the accompanying comments.

The present study draws on several themes in the RV literature. The volatility ratio plots in Figure 2 serve a purpose similar to the volatility signature plots introduced by Fang (1996) and used in Andersen, Bollerslev, Diebold and Ebens (2002) and Hansen and Lunde (2006). Hansen and Lunde also articulate the connection between bid-offer comovement and fundamental volatility: since the bid and offer have economic fundamentals in common, divergent movements must be short-term, transient, and unconnected to fundamentals.

The paper also departs from the RV literature in significant respects. The millisecond time scales employed in this paper are several orders of magnitude shorter than those typically encountered. Most RV studies also focus on relatively liquid assets (index securities, Dow-Jones stocks, etc.). The low-activity securities included in the present paper's samples are important because, due to their larger spreads and fewer participants, they are likely to exhibit relatively strong, persistent and distinctive microstructure-related components.

VIII. Conclusion and outstanding questions

High-frequency volatility in the bid and offer quotes induces risk for agents who experience delay in communicating with the market. The risk may be quantified as the price variance over the interval of delay, relative to the time-weighted average price (TWAP) over the interval. This volatility degrades the informational value of the quotes. Furthermore, because the bid and offer are often used as reference prices for dealer trades against customers, the volatility increases the value of a dealer's look-back option and exacerbates the customer's monitoring problem.

This study is a preliminary analysis of short-term in the US equity market. Applying standard techniques of time-decomposition to a recent sample of millisecond-stamped data establishes that there is substantial volatility in the National Best Bid and Offer (NBBO) on millisecond-level time scales that is well in excess of what would be expected using random-walk volatility estimated over longer intervals. The excess volatility is more pronounced for stocks that have lower average activity. Furthermore, the correlations between bids and offers at these time scales are positive, but low. That the bid and offer are not moving together also suggests that the volatility is not fundamental.

The paper proposes a Bayesian simulation approach to measuring millisecond-level volatility in US equity data (like the Monthly TAQ) that possess all quote records, but are time-stamped only to the second. The approach is validated in a set of millisecond-stamped data by comparing two sets of estimates: one set based on the original time-stamps; the other based on simulated time stamps.

Using the Bayesian approach, the paper turns to a longer US historical sample, 2001-2011 Monthly TAQ data. Despite the current public scrutiny of high-frequency trading, the large growth in the number of quote records, and the presumption that low-latency technology is a new and recent phenomenon, the data suggest at best only a modest rise in short-term quote volatility.

References

Amihud, Yakov, and Haim Mendelson, 1987, Trading mechanisms and stock returns: An empirical investigation, *Journal of Finance* 42, 533-553.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens, 2001, The distribution of realized stock return volatility, *Journal of Financial Economics* 61, 43-76.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys, 2003a, The distribution of realized exchange rate volatility (vol 96, pg 43, 2001), *Journal of the American Statistical Association* 98, 501-501.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys, 2003b, Modeling and forecasting realized volatility, *Econometrica* 71, 579-625.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and H. Ebens, 2002, Great realizations, *Risk* 13, 105-108.

Atsalakis, George S., and Kimon P. Valavanis, 2009, Surveying stock market forecasting techniques - part ii: Soft computing methods, *Expert Systems with Applications* 36, 5932-5941.

Barnea, Amir, 1974, Performance evaluation of new york stock exchange specialists, *Journal of Financial and Quantitative Analysis* 9, 511-535.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2012, Equilibrium high-frequency trading, *SSRN eLibrary*.

Brogaard, Jonathan, 2010, High frequency trading and its impact on market quality, *SSRN eLibrary*.

Brogaard, Jonathan, 2012, High frequency trading and volatility, *SSRN eLibrary*.

Brogaard, Jonathan, Terrence J. Hendershott, and Ryan Riordan, 2012, High frequency trading and price discovery, *SSRN eLibrary*.

Charles, Amélie, and Olivier Darné, 2009, Variance-ratio tests of random walk: An overview, *Journal of Economic Surveys* 23, 503-527.

Dacorogna, Michel M., Ramazan Gencay, Ulrich A. Muller, Richard B. Olsen, and Olivier B. Pictet, 2001. *High-frequency finance* (Academic Press, New York).

Duffie, Darrell, 2010, Presidential address: Asset price dynamics with slow-moving capital, *The Journal of finance* 65, 1237-1267.

Elder, John, and Hyun J. Jin, 2007, Long memory in commodity futures volatility: A wavelet perspective, *Journal of Futures Markets* 27, 411-437.

Elkins-McSherry, 2012, Methodology, https://www.elkinsmcsherry.com/EM/methodology.html, accessed on November 2, 2012.

Engle, Robert F., and Jeffrey R. Russell, 1998, Autoregressive conditional duration: A new model for irregularly spaced transaction data, *Econometrica* 66, 1127-1162.

Engle, Robert, Robert Ferstenberg, and Jeffrey Russell, 2012, Measuring and modeling execution cost and risk, *Journal of Portfolio Management* 38, 14-28.

Fan, Jianqing, and Yazhen Wang, 2007, Multi-scale jump and volatility analysis for high-frequency financial data, *Journal of the American Statistical Association* 102, 1349-1362.

Fang, Y., 1996, Volatility modeling and estimation of high-frequency data with gaussian noise, (MIT Sloan School).

Faust, Jon, 1992, When are variance ratio tests for serial dependence optimal?, *Econometrica* 60, 1215-1226.

Gençay, Ramazan, Faruk Selçuk, Nikola Gradojevic, and Brandon Whitcher, 2009, Asymmetry of information flow between volatilities across time scales, (SSRN).

Gençay, Ramazan, Frank Selçuk, and Brandon Whitcher, 2002. *An introduction to wavelets and other filtering methods in finance and economics* (Academic Press (Elsevier), San Diego).

Hansen, Peter R., and Asger Lunde, 2006, Realized variance and market microstructure noise, *Journal of Business & Economic Statistics* 24, 127-161.

Harris, Lawrence E., 1999, Trading in pennies: A survey of the issues, (Marshall School, University of Southern California).

Hasbrouck, Joel, and Gideon Saar, 2011, Low-latency trading, (SSRN eLibrary).

Hasbrouck, Joel, and Robert A. Schwartz, 1988, Liquidity and execution costs in equity markets, *Journal of Portfolio Management* 14, 10-16.

Hautsch, Nikolaus, 2004. *Modelling irregularly spaced financial data: Theory and practice of dynamic duration models* (Springer).

Hendershott, Terrence J., and Ryan Riordan, 2012, Algorithmic trading and the market for liquidity, *SSRN eLibrary*.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld, 2011, Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1-33.

Høg, Esben, and Asger Lunde, 2003, Wavelet estimation of integrated volatility, (Aarhus School of Business, Department of Information Science).

Holden, Craig W., and Stacey E. Jacobsen, 2012, Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions, (Kelley School, University of Indiana).

Hsieh, Tsung-Jung, Hsiao-Fen Hsiao, and Wei-Chang Yeh, 2011, Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm, *Applied Soft Computing* 11, 2510-2525.

In, Francis, and Sangbae Kim, 2006, The hedge ratio and the empirical relationship between the stock and futures markets: A new approach using wavelet analysis, *The Journal of Business* 79, 799-820.

Jarrow, Robert A., and Philip Protter, 2011, A dysfunctional role of high frequency trading in electronic markets, *SSRN eLibrary*.

Jennings, Robert H., and Ramabhadran S. Thirumalai, 2007, Advertising for liquidity on the new york stock exchange, (Kelley School, University of Indiana).

Kirilenko, Andrei A., Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun, 2010, The flash crash: The impact of high frequency trading on an electronic market, *SSRN eLibrary*.

Lo, Andrew W., and A. Craig MacKinlay, 1989, The size and power of the variance ratio test in finite samples, *Journal of Econometrics* 40, 203-238.

Menkveld, Albert J., 2012, High frequency trading and the new market makers, *SSRN eLibrary*.

Nason, Guy P., 2008. *Wavelet methods in statistics with r* (Springer Science+Business Media, LLC, New York).

Pagnotta, Emiliano S., 2009, Trading strategies at optimal frequencies, (Stern School).

Parlour, Christine A., and Duane Seppi, 2008, Limit order markets: A survey, in Anjan V. Thakor, and Arnaud W. A. Boot, eds.: *Handbook of financial intermediation and banking* (Elsevier, Amsterdam).

Parlour, Christine A., and Duane J. Seppi, 2003, Liquidity-based competition for order flow, *Review of financial studies* 16, 301-303.

Percival, Donald B., and Andrew T. Walden, 2000. *Wavelet methods for time series analysis* (Cambridge University Press, Cambridge).

Perold, Andre, 1988, The implementation shortfall: Paper vs. Reality, *Journal of Portfolio Management* 14, 4-9.

Ramsey, James B., 1999, The contribution of wavelets to the analysis of economic and financial data, *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 357, 2593-2606.

Ramsey, James B., 2002, Wavelets in economics and finance: Past and future, *Studies in Nonlinear Dynamics and Econometrics* 6.

Ross, Sheldon M., 1996. *Stochastic processes* (John Wiley and Sons, Inc., New York).

Shephard, Neil, 2005, General introduction, in Neil Shephard, ed.: *Stochastic volatility* (Oxford University Press, Oxford).

Stoll, H. R., and C. Schenzler, 2006, Trades outside the quotes: Reporting delay, trading option, or trade size?, *Journal of Financial Economics* 79, 615-653.

Teyssière, Gilles, and Patrice Abry, 2007, Wavelet analysis of nonlinear long-range dependent processes: Applications to financial time series, in Gilles Teyssière, and Alan P. Kirman, eds.: *Long memory in economics* (Springer, Berlin).

U.S. Commodities Futures Trading Commission, 2011, Presentation (working group 1, additional document), Technology Advisor Committee.

U.S. Commodities Futures Trading Commission Technology Advisor Committee, 2001, Market access committee interim report.

U.S. Securities and Exchange Commission, 2010, Concept release on equity market structure.

U.S. Securities and Exchange Commission, 2012, Consolidated audit trail (rule 613, final rule).

Walker, James S., 2008. *A primer on wavelets and their scientific applications* (Chapman and Hall/CRC (Taylor and Francis Group), Boca Raton).

Table 1. Sample Summary Statistics

Source: CRSP and Daily TAQ data, April 2011. The sample is 100 firms randomly selected from CRSP with stratification based on average dollar trading volume in the first quarter of 2011, grouped in quintiles by dollar trading volume over the first quarter of 2011.

			Cross-firm median					
		N	Share price, end of 2010	Avg. Dollar Volume ($ Thousand)	Equity Market Capitalization ($Million)	Avg. no. daily trades	Avg. no. daily quote updates	Avg. no. daily NBBO records
Full sample		100	$13.93	$2,140	$399	1,061	23,347	6,897
Dollar volume quintiles	1, low	20	$4.18	$39	$30	30	960	362
	2	20	$4.37	$501	$148	404	6,288	2,768
	3	20	$9.54	$2,140	$372	873	20,849	6,563
	4	20	$27.82	$9,691	$1,440	2,637	45,728	11,512
	5, high	20	$39.33	$85,324	$6,231	13,057	179,786	37,695

Table 2. Volatility in the National Best Bid and Offer

The results are based on a random sample of 100 US stocks, stratified by dollar trading volume, for April, 2011. Wavelet variance estimates by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer at one-millisecond resolution. They are averaged over the bid and offer sides, and cumulated to obtain estimates of wavelet rough variances, $Var(\mathcal{R}_j)$. Table entries are cross-firm sample means and (in parentheses) standard errors for $\sqrt{Var(\mathcal{R}_j)}$ in mils per share ($0.001, Panel A) and for $\sqrt{Var(\mathcal{R}_j)}/\bar{p}$, where $\bar{p}$ is the firm's average bid-ask midpoint over the sample, in basis points (0.01%, Panel B).

Panel A. $\sqrt{Var(\mathcal{R}_j)}$ at time scales τ_j, mils ($0.001) per share.

			Dollar volume quintiles									
Time scale	Full Sample		1 (low)		2		3		4		5 (high)	
64 ms	0.4	(<0.1)	0.3	(<0.1)	0.2	(<0.1)	0.4	(<0.1)	0.5	(<0.1)	0.6	(0.1)
128 ms	0.6	(<0.1)	0.4	(<0.1)	0.3	(<0.1)	0.6	(<0.1)	0.7	(<0.1)	0.9	(0.2)
256 ms	0.8	(<0.1)	0.6	(0.1)	0.5	(<0.1)	0.8	(0.1)	1.0	(0.1)	1.2	(0.3)
512 ms	1.1	(0.1)	0.8	(0.2)	0.6	(<0.1)	1.2	(0.2)	1.4	(0.2)	1.7	(0.4)
1,024 ms	1.6	(0.1)	1.1	(0.2)	0.9	(0.1)	1.6	(0.3)	1.9	(0.3)	2.4	(0.6)
4.1 sec	3.0	(0.3)	2.1	(0.4)	1.6	(0.3)	3.0	(0.5)	3.6	(0.5)	4.8	(1.1)
32.8 sec	8.0	(0.8)	5.3	(1.0)	4.1	(0.7)	7.4	(1.1)	9.9	(1.5)	13.3	(3.1)

Panel B. $\sqrt{Var(\mathcal{R}_j)}/\bar{p}$, basis points (0.01%)

			Dollar volume quintiles									
Time scale	Full Sample		1 (low)		2		3		4		5 (high)	
64 ms	0.3	(0.02)	0.6	(0.08)	0.4	(0.04)	0.3	(0.02)	0.2	(0.01)	0.1	(0.01)
128 ms	0.4	(0.03)	0.8	(0.11)	0.6	(0.05)	0.4	(0.03)	0.2	(0.02)	0.2	(0.01)
256 ms	0.6	(0.04)	1.1	(0.15)	0.8	(0.07)	0.6	(0.05)	0.3	(0.02)	0.2	(0.02)
512 ms	0.8	(0.06)	1.4	(0.20)	1.1	(0.10)	0.8	(0.07)	0.5	(0.03)	0.3	(0.02)
1,024 ms	1.2	(0.08)	2.0	(0.27)	1.5	(0.14)	1.2	(0.10)	0.6	(0.05)	0.5	(0.04)
4.1 sec	2.2	(0.16)	3.8	(0.52)	2.8	(0.27)	2.2	(0.21)	1.2	(0.11)	0.9	(0.08)
32.8 sec	5.6	(0.40)	9.5	(1.26)	7.1	(0.75)	5.7	(0.56)	3.3	(0.34)	2.5	(0.24)

Table 3. Wavelet variance ratios for the National Best Bid and Offer

The results are based on a random sample of 100 US stocks, stratified by dollar trading volume, for April, 2011. Wavelet variance estimates $\nu^2(\tau_j)$ by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer at one-millisecond resolution, and (for time scales of a second or longer) at a one-second resolution. " Base res" indicates the resolution in the $\nu^2(\tau_j)$ calculation. The estimates are averaged across the NBB and NBO. The wavelet variance ratio is

$$VR(\tau_j) = \frac{\nu^2(\tau_j)}{\nu^2(\tau_J)} \times \frac{\nu^2_{\overline{rw}}(\tau_J)}{\nu^2_{\overline{rw}}(\tau_j)}$$

$\nu^2(\tau_J)$ is the wavelet variance at the longest time scale in the analysis (34.1 minutes). The normalization ratio $\nu^2_{\overline{rw}}(\tau_J)/\nu^2_{rw}(\tau_j)$ is based on the assumption that the prices follow random-walks at a one millisecond resolution. The "$\overline{rw}$" subscript indicates that the variance has also been adjusted for averaging at the one-second resolution. Due to the normalization, $VR(\tau_j = 34.1\ min)$ is unity by construction. Table entries are cross-firm means and (in parentheses) the standard errors of these means.

				Dollar volume quintiles									
Time scale τ_j	Base Res	Full Sample		1		2		3		4		5	
1 ms	ms	5.36	(0.93)	8.81	(2.37)	9.74	(3.71)	3.69	(0.44)	2.74	(0.36)	1.83	(0.17)
2 ms	ms	5.21	(0.92)	8.56	(2.32)	9.73	(3.71)	3.66	(0.45)	2.40	(0.24)	1.70	(0.14)
4 ms	ms	4.73	(0.82)	7.26	(1.76)	9.01	(3.46)	3.52	(0.44)	2.26	(0.23)	1.59	(0.12)
8 ms	ms	4.18	(0.71)	5.86	(1.31)	8.05	(3.15)	3.33	(0.41)	2.12	(0.21)	1.53	(0.11)
16 ms	ms	3.79	(0.65)	5.11	(1.15)	7.26	(2.91)	3.08	(0.37)	2.00	(0.20)	1.49	(0.11)
32 ms	ms	3.48	(0.59)	4.62	(1.03)	6.55	(2.63)	2.85	(0.32)	1.92	(0.19)	1.48	(0.10)
64 ms	ms	3.25	(0.52)	4.31	(0.94)	5.93	(2.33)	2.69	(0.29)	1.86	(0.18)	1.47	(0.10)
128 ms	ms	3.03	(0.46)	4.03	(0.88)	5.35	(2.04)	2.56	(0.27)	1.78	(0.17)	1.45	(0.10)
256 ms	ms	2.82	(0.41)	3.75	(0.85)	4.85	(1.79)	2.41	(0.24)	1.69	(0.16)	1.40	(0.09)
512 ms	ms	2.69	(0.38)	3.71	(0.86)	4.52	(1.63)	2.30	(0.22)	1.58	(0.14)	1.34	(0.09)
1,000 ms	sec	2.56	(0.34)	3.63	(0.85)	4.13	(1.42)	2.19	(0.20)	1.51	(0.13)	1.31	(0.09)
1,024 ms	ms	2.54	(0.35)	3.63	(0.84)	4.13	(1.44)	2.17	(0.20)	1.49	(0.13)	1.30	(0.09)
2,000 ms	sec	2.37	(0.29)	3.45	(0.79)	3.64	(1.16)	2.05	(0.18)	1.43	(0.12)	1.27	(0.08)
2,048 ms	ms	2.36	(0.29)	3.45	(0.79)	3.63	(1.16)	2.04	(0.17)	1.42	(0.12)	1.27	(0.08)
4.0 sec	sec	2.18	(0.23)	3.21	(0.68)	3.15	(0.88)	1.90	(0.15)	1.36	(0.10)	1.25	(0.08)
4.1 sec	ms	2.17	(0.23)	3.20	(0.68)	3.14	(0.87)	1.90	(0.15)	1.36	(0.10)	1.25	(0.08)
8.0 sec	sec	1.97	(0.18)	2.87	(0.53)	2.69	(0.63)	1.75	(0.13)	1.30	(0.09)	1.25	(0.08)
8.2 sec	ms	1.97	(0.17)	2.87	(0.52)	2.68	(0.62)	1.75	(0.13)	1.30	(0.09)	1.25	(0.08)
16.0 sec	sec	1.75	(0.12)	2.39	(0.34)	2.29	(0.44)	1.59	(0.11)	1.24	(0.08)	1.24	(0.08)
16.4 sec	ms	1.75	(0.12)	2.38	(0.34)	2.29	(0.43)	1.59	(0.11)	1.24	(0.08)	1.25	(0.08)
32.0 sec	sec	1.60	(0.09)	2.13	(0.27)	1.99	(0.30)	1.47	(0.10)	1.18	(0.07)	1.22	(0.08)
32.8 sec	ms	1.60	(0.09)	2.13	(0.26)	1.99	(0.30)	1.47	(0.10)	1.19	(0.07)	1.23	(0.08)
64.0 sec	sec	1.46	(0.06)	1.91	(0.20)	1.72	(0.20)	1.35	(0.09)	1.14	(0.07)	1.18	(0.06)
2.1 min	sec	1.34	(0.05)	1.71	(0.15)	1.49	(0.13)	1.26	(0.08)	1.10	(0.06)	1.13	(0.05)
4.3 min	sec	1.24	(0.03)	1.54	(0.12)	1.33	(0.08)	1.18	(0.07)	1.08	(0.05)	1.09	(0.04)
8.5 min	sec	1.16	(0.02)	1.34	(0.07)	1.21	(0.05)	1.09	(0.04)	1.07	(0.04)	1.08	(0.03)
17.1 min	sec	1.08	(0.01)	1.15	(0.03)	1.12	(0.03)	1.02	(0.02)	1.04	(0.02)	1.06	(0.02)
34.1 min	sec	1.00	(0.00)	1.00	(0.00)	1.00	(0.00)	1.00	(0.00)	1.00	(0.00)	1.00	(0.00)

Table 4. Wavelet correlations between the National Best Bid and National Best Offer.

The results are based on a random sample of 100 US stocks, stratified by dollar trading volume, for April, 2011. Wavelet variance and covariance estimates by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer at one-millisecond resolution, and (for time scales of a second or longer) at a one-second resolution. "Base resolution" indicates the resolution in the calculation. Denoting these wavelet variances and covariances by $\nu_{NBB}^2(\tau_j)$, $\nu_{NBO}^2(\tau_j)$, $\nu_{NBB,NBO}(\tau_j)$, the wavelet correlated is defined as $\nu_{NBB,NBO}(\tau_j)/\sqrt{\nu_{NBB}^2(\tau_j) \times \nu_{NBO}^2(\tau_j)}$. Table entries are cross-firm medians.

			Dollar volume quintile				
Time scale	Base Resolution	Full Sample	1 (low)	2	3	4	5 (high)
1 ms	ms	0.068	0.007	0.051	0.066	0.079	0.135
2 ms	ms	0.107	0.015	0.058	0.086	0.130	0.245
4 ms	ms	0.156	0.024	0.082	0.125	0.195	0.357
8 ms	ms	0.207	0.031	0.117	0.175	0.257	0.454
16 ms	ms	0.253	0.044	0.155	0.224	0.307	0.536
32 ms	ms	0.295	0.062	0.195	0.271	0.349	0.596
64 ms	ms	0.335	0.083	0.238	0.319	0.391	0.641
128 ms	ms	0.375	0.109	0.289	0.370	0.430	0.678
256 ms	ms	0.414	0.139	0.340	0.417	0.467	0.708
512 ms	ms	0.447	0.165	0.377	0.454	0.503	0.736
1,024 ms	ms	0.481	0.183	0.412	0.490	0.547	0.771
2,048 ms	ms	0.517	0.199	0.449	0.528	0.599	0.811
4.1 sec	ms	0.555	0.211	0.484	0.572	0.656	0.855
8.2 sec	ms	0.598	0.230	0.523	0.623	0.719	0.897
16.4 sec	ms	0.644	0.256	0.570	0.678	0.781	0.933
32.8 sec	ms	0.689	0.284	0.622	0.737	0.842	0.960
64.0 sec	sec	0.735	0.319	0.685	0.800	0.896	0.977
2.1 min	sec	0.779	0.352	0.750	0.863	0.940	0.988
4.3 min	sec	0.814	0.383	0.805	0.916	0.970	0.995
8.5 min	sec	0.842	0.424	0.849	0.953	0.987	0.998
17.1 min	sec	0.864	0.469	0.884	0.974	0.994	0.999
34.1 min	sec	0.882	0.520	0.908	0.986	0.997	1.000

Table 5. Agreement of wavelet variances and correlations constructed using original and simulated time stamps.

The results are based on a random sample of 100 US stocks, stratified by dollar trading volume, for all trading days in April, 2011. Wavelet variance and correlation estimates by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer over 15 minute intervals from 9:45 to 15:45. For each firm, time scale and interval, the wavelet variances are averaged across the NBB and NBO. These estimates are formed twice: using in the first instance using the actual millisecond time stamps, and in the second instance using simulated millisecond time stamps. For each firm I compute the correlation between corresponding actual/simulated estimates. The table reports the cross-firm medians for these correlations.

Panel A. Wavelet variances

		Dollar volume quintiles				
Time Scale	All	1 (low)	2	3	4	5 (high)
1 ms	0.991	0.997	0.985	0.991	0.995	0.989
2 ms	0.986	0.990	0.985	0.989	0.990	0.981
4 ms	0.978	0.979	0.976	0.978	0.980	0.980
8 ms	0.973	0.967	0.959	0.968	0.977	0.979
16 ms	0.967	0.947	0.959	0.965	0.975	0.979
32 ms	0.966	0.929	0.946	0.968	0.974	0.981
64 ms	0.968	0.920	0.952	0.963	0.978	0.985
128 ms	0.970	0.929	0.955	0.961	0.978	0.987
256 ms	0.975	0.960	0.962	0.973	0.984	0.990
512 ms	0.986	0.981	0.979	0.983	0.991	0.995
1.0 sec	0.996	0.992	0.993	0.995	0.997	0.998
2.0 sec	0.999	0.998	0.998	0.998	0.999	1.000
4.1 sec	1.000	1.000	0.999	0.999	1.000	1.000
8.2 sec	1.000	1.000	1.000	1.000	1.000	1.000
16.4 sec	1.000	1.000	1.000	1.000	1.000	1.000
32.8 sec	1.000	1.000	1.000	1.000	1.000	1.000

Panel B. Wavelet correlations

		Dollar volume quintiles				
Time Scale	All	1 (low)	2	3	4	5 (high)
1 ms	0.219	0.304	0.136	0.247	0.221	0.226
2 ms	0.247	0.198	0.168	0.276	0.243	0.352
4 ms	0.289	0.255	0.210	0.313	0.278	0.393
8 ms	0.330	0.218	0.342	0.329	0.316	0.436
16 ms	0.377	0.261	0.376	0.384	0.393	0.462
32 ms	0.424	0.282	0.404	0.401	0.482	0.486
64 ms	0.503	0.388	0.464	0.491	0.565	0.616
128 ms	0.594	0.488	0.545	0.612	0.671	0.737
256 ms	0.715	0.611	0.666	0.741	0.784	0.823
512 ms	0.862	0.803	0.816	0.865	0.887	0.906
1.0 sec	0.949	0.946	0.934	0.943	0.957	0.965
2.0 sec	0.984	0.982	0.980	0.982	0.987	0.989
4.1 sec	0.995	0.994	0.994	0.994	0.996	0.995
8.2 sec	0.998	0.998	0.998	0.998	0.999	0.998
16.4 sec	0.999	0.999	0.999	0.999	0.999	0.999
32.8 sec	0.999	1.000	0.999	0.999	1.000	0.999

Table 6. Summary statistics, historical sample, 2001-2011

For each year, 2001-2011 I draw thirty firms randomly, stratified by equity market capitalization as of the end of March. Trade and quote counts are from the Monthly TAQ database (one-second time stamps). Except for the number of firms, table entries are cross-firm medians.

	Year										
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
No. of firms	30	29	30	30	30	30	30	30	30	30	30
Equity mkt. cap., $ Million	$341	$151	$165	$287	$272	$374	$445	$329	$150	$379	$484
Price per share	$15.88	$10.14	$9.45	$15.27	$22.51	$15.26	$16.65	$9.16	$4.90	$8.84	$16.23
Avg. daily no. of trades	97	85	65	347	276	361	889	1,172	869	1,510	1,341
Avg. daily no. of quotes	807	271	814	3,588	4,846	5,761	12,383	14,427	18,305	19,320	17,989

Table 7. Volatility in the National Best Bid and Offer, historical

The sample consists of thirty firms in each year 2001-2011, randomly chosen with equity market capitalization stratification. Quote data are from the Monthly TAQ. The one-second time-stamps on the Monthly TAQ are supplemented with randomized millisecond remainders. Then wavelet variance estimates by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer at one-millisecond resolution. They are averaged over the bid and offer sides, and cumulated to obtain estimates of wavelet rough variances, $Var(\mathcal{R}_j)$. Table entries are cross-firm sample means and (in parentheses) standard errors for $\sqrt{Var(\mathcal{R}_j)}$ in mils per share ($0.001, Panel A) and for $\sqrt{Var(\mathcal{R}_j)}/\bar{p}$, where $\bar{p}$ is the firm's average bid-ask midpoint over the sample, in basis points (0.01%, Panel B).

Panel A. $\sqrt{Var(\mathcal{R}_j)}$ at time scales τ_j, mils ($0.001) per share.

											Year
Time scale	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
64 ms	0.5	0.4	0.4	0.5	0.7	0.4	0.4	0.5	0.5	0.4	0.3
	(<0.1)	(<0.1)	(<0.1)	(<0.1)	(0.1)	(<0.1)	(<0.1)	(<0.1)	(<0.1)	(<0.1)	(<0.1)
128 ms	0.7	0.5	0.5	0.7	1.0	0.6	0.5	0.7	0.7	0.6	0.5
	(0.1)	(0.1)	(<0.1)	(<0.1)	(0.1)	(<0.1)	(<0.1)	(<0.1)	(<0.1)	(0.1)	(<0.1)
256 ms	0.9	0.7	0.7	1.0	1.3	0.8	0.7	0.9	0.9	0.8	0.6
	(0.2)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(<0.1)	(0.1)	(0.1)	(0.1)	(<0.1)
512 ms	1.3	1.0	1.0	1.4	1.9	1.2	0.9	1.2	1.2	1.1	0.9
	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.1)	(0.1)	(0.2)	(0.2)	(0.2)	(0.1)
1,024 ms	1.9	1.4	1.4	1.9	2.5	1.6	1.3	1.7	1.6	1.5	1.2
	(0.3)	(0.3)	(0.3)	(0.2)	(0.3)	(0.2)	(0.1)	(0.2)	(0.2)	(0.3)	(0.1)
4.1 sec	3.6	2.7	2.8	3.5	4.6	2.9	2.3	3.2	3.0	2.7	2.2
	(0.6)	(0.6)	(0.5)	(0.4)	(0.6)	(0.4)	(0.3)	(0.4)	(0.4)	(0.5)	(0.3)
32.8 sec	10.1	7.0	7.2	8.6	11.3	7.4	6.3	8.3	7.7	6.9	5.8
	(1.7)	(1.5)	(1.3)	(1.0)	(1.3)	(0.9)	(0.7)	(1.2)	(1.2)	(1.2)	(0.8)

Panel B. $\sqrt{Var(\mathcal{R}_j)}/\bar{p}$, basis points (0.01%)

											Year
Time scale	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
64 ms	0.3	0.5	0.3	0.5	0.4	0.3	0.2	0.5	1.1	0.4	0.3
	(0.04)	(0.13)	(0.03)	(0.06)	(0.05)	(0.04)	(0.03)	(0.07)	(0.23)	(0.05)	(0.04)
128 ms	0.4	0.7	0.4	0.7	0.6	0.5	0.3	0.7	1.5	0.6	0.4
	(0.06)	(0.18)	(0.05)	(0.09)	(0.08)	(0.06)	(0.05)	(0.10)	(0.32)	(0.07)	(0.06)
256 ms	0.6	1.0	0.6	0.9	0.8	0.6	0.4	0.9	2.0	0.8	0.5
	(0.08)	(0.26)	(0.07)	(0.13)	(0.11)	(0.09)	(0.06)	(0.13)	(0.45)	(0.10)	(0.08)
512 ms	0.8	1.3	0.8	1.3	1.2	0.9	0.6	1.3	2.7	1.1	0.7
	(0.12)	(0.37)	(0.11)	(0.18)	(0.15)	(0.12)	(0.09)	(0.18)	(0.62)	(0.14)	(0.11)
1,024 ms	1.1	1.8	1.2	1.7	1.6	1.2	0.8	1.7	3.7	1.4	1.0
	(0.17)	(0.50)	(0.15)	(0.25)	(0.20)	(0.16)	(0.12)	(0.23)	(0.84)	(0.20)	(0.16)
4.1 sec	2.2	3.7	2.2	3.2	2.9	2.1	1.5	3.1	6.6	2.6	1.9
	(0.35)	(1.03)	(0.30)	(0.44)	(0.33)	(0.27)	(0.22)	(0.40)	(1.49)	(0.38)	(0.29)
32.8 sec	6.2	9.3	5.6	7.7	6.9	5.2	3.8	7.5	14.6	6.3	4.7
	(0.98)	(2.51)	(0.70)	(1.06)	(0.74)	(0.60)	(0.54)	(0.83)	(2.60)	(0.98)	(0.70)

Table 8. Wavelet variance ratios for the National Best Bid and Offer, historical

The sample consists of thirty firms in each year 2001-2011, randomly chosen with equity market capitalization stratification. Quote data are from the Monthly TAQ. The one-second time-stamps on the Monthly TAQ are supplemented with randomized millisecond remainders. Wavelet variance estimates $\nu^2(\tau_j)$ by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer at one-millisecond resolution, and (for time scales of a second or longer) at a one-second resolution. " Base res" indicates the resolution in the $\nu^2(\tau_j)$ calculation. The estimates are averaged across the NBB and NBO. The wavelet variance ratio is

$$VR(\tau_j) = \frac{\nu^2(\tau_j)}{\nu^2(\tau_J)} \times \frac{\nu^2_{\overline{rw}}(\tau_J)}{\nu^2_{rw}(\tau_j)}$$

$\nu^2(\tau_J)$ is the wavelet variance at the longest time scale in the analysis (34.1 minutes). The normalization ratio $\nu^2_{\overline{rw}}(\tau_J)/\nu^2_{rw}(\tau_j)$ is based on the assumption that the prices follow random-walks at a one millisecond resolution. The "$\overline{rw}$" subscript indicates that the variance has also been adjusted for averaging at the one-second resolution. Due to the normalization, $VR(\tau_j = 34.1\ min)$ is unity by construction. Table entries are cross-firm means and (in parentheses) the standard errors of these means.

Time scale	Base res	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
1ms	ms	2.41	2.45	4.99	7.63	6.89	4.69	4.86	5.76	7.79	9.30	5.97
		(0.46)	(0.20)	(1.23)	(1.80)	(1.76)	(0.80)	(1.44)	(1.49)	(1.67)	(2.86)	(1.36)
2 ms	ms	2.41	2.45	4.99	7.62	6.88	4.66	4.84	5.72	7.76	9.17	5.90
		(0.46)	(0.20)	(1.23)	(1.80)	(1.76)	(0.78)	(1.43)	(1.49)	(1.67)	(2.84)	(1.34)
4 ms	ms	2.41	2.45	4.98	7.61	6.88	4.62	4.80	5.66	7.72	8.98	5.79
		(0.46)	(0.20)	(1.23)	(1.80)	(1.76)	(0.76)	(1.41)	(1.48)	(1.67)	(2.81)	(1.31)
8 ms	ms	2.41	2.44	4.98	7.60	6.86	4.53	4.72	5.56	7.61	8.70	5.62
		(0.46)	(0.20)	(1.23)	(1.80)	(1.75)	(0.72)	(1.37)	(1.46)	(1.66)	(2.76)	(1.27)
16 ms	ms	2.40	2.41	4.97	7.58	6.83	4.41	4.59	5.38	7.43	8.29	5.36
		(0.46)	(0.20)	(1.23)	(1.80)	(1.75)	(0.66)	(1.31)	(1.42)	(1.63)	(2.70)	(1.19)
32 ms	ms	2.39	2.38	4.96	7.53	6.78	4.28	4.37	5.11	7.17	7.78	4.96
		(0.46)	(0.19)	(1.22)	(1.79)	(1.73)	(0.62)	(1.21)	(1.36)	(1.59)	(2.63)	(1.06)
64 ms	ms	2.37	2.35	4.92	7.44	6.66	4.13	4.06	4.73	6.83	7.23	4.51
		(0.45)	(0.19)	(1.21)	(1.77)	(1.70)	(0.59)	(1.07)	(1.27)	(1.55)	(2.57)	(0.93)
128 ms	ms	2.34	2.31	4.83	7.26	6.45	3.94	3.67	4.26	6.39	6.69	4.03
		(0.45)	(0.18)	(1.19)	(1.74)	(1.62)	(0.54)	(0.90)	(1.13)	(1.48)	(2.53)	(0.78)
256 ms	ms	2.31	2.25	4.67	6.95	6.06	3.70	3.22	3.74	5.84	6.16	3.52
		(0.45)	(0.18)	(1.13)	(1.69)	(1.48)	(0.49)	(0.74)	(0.95)	(1.39)	(2.50)	(0.63)
512 ms	ms	2.27	2.18	4.42	6.46	5.46	3.39	2.76	3.23	5.22	5.62	3.03
		(0.45)	(0.17)	(1.07)	(1.61)	(1.24)	(0.44)	(0.59)	(0.75)	(1.27)	(2.47)	(0.50)
1,024 ms	ms	2.20	2.11	4.13	5.83	4.66	3.06	2.40	2.83	4.60	5.13	2.65
		(0.42)	(0.16)	(1.04)	(1.53)	(0.92)	(0.37)	(0.50)	(0.61)	(1.13)	(2.42)	(0.40)
2,048 ms	ms	2.18	2.05	3.87	5.29	3.92	2.76	2.18	2.56	4.13	4.75	2.39
		(0.41)	(0.16)	(1.04)	(1.50)	(0.66)	(0.32)	(0.44)	(0.50)	(1.00)	(2.40)	(0.33)
4.1 sec	ms	2.22	1.99	3.67	4.92	3.48	2.50	2.07	2.35	3.68	4.46	2.18
		(0.44)	(0.15)	(1.04)	(1.49)	(0.54)	(0.26)	(0.41)	(0.43)	(0.81)	(2.39)	(0.27)
8.2 sec	ms	2.23	1.91	3.50	4.56	3.13	2.27	1.98	2.16	3.10	4.26	1.99
		(0.45)	(0.14)	(1.03)	(1.51)	(0.45)	(0.22)	(0.38)	(0.37)	(0.56)	(2.40)	(0.22)
16.4 sec	ms	2.17	1.78	3.27	4.21	2.82	2.07	1.75	1.96	2.59	4.10	1.80
		(0.45)	(0.12)	(1.01)	(1.57)	(0.38)	(0.18)	(0.23)	(0.31)	(0.36)	(2.40)	(0.17)
32.8 sec	ms	2.03	1.64	2.80	3.99	2.53	1.90	1.58	1.77	2.20	3.96	1.65
		(0.44)	(0.11)	(0.85)	(1.65)	(0.31)	(0.15)	(0.14)	(0.26)	(0.25)	(2.37)	(0.13)
64.0 sec	Sec	1.89	1.52	2.13	3.87	2.25	1.74	1.47	1.60	2.07	2.83	1.54
		(0.45)	(0.09)	(0.40)	(1.79)	(0.25)	(0.12)	(0.09)	(0.20)	(0.26)	(1.34)	(0.10)
2.1 min	Sec	1.83	1.42	1.71	3.98	1.97	1.59	1.40	1.44	1.94	2.12	1.44
		(0.47)	(0.08)	(0.14)	(2.15)	(0.19)	(0.09)	(0.07)	(0.13)	(0.25)	(0.71)	(0.08)

Time scale	Base res	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
4.3 min	Sec	1.75	1.34	1.49	3.60	1.72	1.47	1.36	1.32	1.83	2.23	1.36
		(0.44)	(0.06)	(0.07)	(1.97)	(0.14)	(0.07)	(0.06)	(0.07)	(0.24)	(0.91)	(0.07)
8.5 min	Sec	1.65	1.20	1.29	1.52	1.42	1.30	1.26	1.18	1.64	1.36	1.20
		(0.45)	(0.04)	(0.04)	(0.13)	(0.08)	(0.04)	(0.05)	(0.04)	(0.22)	(0.20)	(0.04)
17.1 min	Sec	1.27	1.10	1.17	1.23	1.20	1.17	1.14	1.08	1.27	1.10	1.11
		(0.17)	(0.02)	(0.03)	(0.04)	(0.03)	(0.02)	(0.03)	(0.02)	(0.08)	(0.04)	(0.02)
34.1 min	Sec	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Table 9. Wavelet correlations, National Best Bid and National Best Offer, historical

The sample consists of thirty firms in each year 2001-2011, randomly chosen with equity market capitalization stratification. Quote data are from the Monthly TAQ. The one-second time-stamps on the Monthly TAQ are supplemented with randomized millisecond remainders. Wavelet variance and covariance estimates by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer at one-millisecond resolution, and (for time scales of a second or longer) at a one-second resolution. "Base resolution" indicates the resolution in the calculation. Denoting these wavelet variances and covariances by $\nu^2_{NBB}(\tau_j)$, $\nu^2_{NBO}(\tau_j)$, $\nu_{NBB,NBO}(\tau_j)$, the wavelet correlated is defined as $\nu_{NBB,NBO}(\tau_j)/\sqrt{\nu^2_{NBB}(\tau_j)\times\nu^2_{NBO}(\tau_j)}$. Table entries are cross-firm medians.

		Year										
Time scale	Base resolution	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
1 ms	ms	0.068	0.038	0.024	0.004	0.052	0.053	0.018	0.018	0.013	0.007	0.005
2 ms	ms	0.068	0.038	0.024	0.004	0.052	0.053	0.018	0.019	0.013	0.009	0.008
4 ms	ms	0.069	0.038	0.024	0.004	0.052	0.054	0.019	0.021	0.016	0.013	0.012
8 ms	ms	0.069	0.038	0.024	0.005	0.053	0.054	0.020	0.026	0.020	0.020	0.020
16 ms	ms	0.069	0.039	0.024	0.005	0.054	0.056	0.022	0.035	0.029	0.033	0.034
32 ms	ms	0.069	0.039	0.024	0.006	0.055	0.058	0.027	0.051	0.045	0.055	0.057
64 ms	ms	0.069	0.039	0.024	0.009	0.059	0.063	0.036	0.078	0.074	0.089	0.093
128 ms	ms	0.070	0.041	0.026	0.013	0.065	0.072	0.054	0.120	0.118	0.142	0.147
256 ms	ms	0.072	0.043	0.030	0.023	0.078	0.090	0.092	0.183	0.179	0.218	0.224
512 ms	ms	0.075	0.048	0.041	0.043	0.101	0.122	0.155	0.268	0.247	0.316	0.323
1,024 ms	ms	0.080	0.062	0.070	0.079	0.139	0.176	0.245	0.356	0.312	0.406	0.414
2,048 ms	ms	0.092	0.093	0.118	0.133	0.191	0.246	0.342	0.426	0.368	0.472	0.477
4.1 sec	ms	0.122	0.138	0.177	0.192	0.250	0.320	0.420	0.480	0.419	0.523	0.522
8.2 sec	ms	0.169	0.193	0.247	0.260	0.318	0.391	0.490	0.530	0.471	0.573	0.564
16.4 sec	ms	0.229	0.262	0.325	0.336	0.397	0.461	0.566	0.584	0.525	0.628	0.607
32.8 sec	ms	0.306	0.347	0.403	0.414	0.483	0.535	0.646	0.636	0.576	0.683	0.649
64.0 sec	sec	0.396	0.436	0.473	0.492	0.569	0.610	0.723	0.687	0.627	0.731	0.689
2.1 min	sec	0.496	0.522	0.545	0.575	0.650	0.683	0.792	0.733	0.675	0.773	0.729
4.3 min	sec	0.588	0.597	0.620	0.653	0.714	0.747	0.842	0.769	0.719	0.808	0.758
8.5 min	sec	0.663	0.660	0.686	0.729	0.770	0.803	0.877	0.797	0.755	0.847	0.781
17.1 min	sec	0.720	0.715	0.740	0.805	0.815	0.833	0.895	0.818	0.787	0.868	0.801
34.1 min	sec	0.762	0.760	0.783	0.851	0.849	0.852	0.912	0.838	0.816	0.882	0.821

Figure 1. The National Best Bid and Offer for AEPI, April 29, 2011

Source: NYSE Daily TAQ data

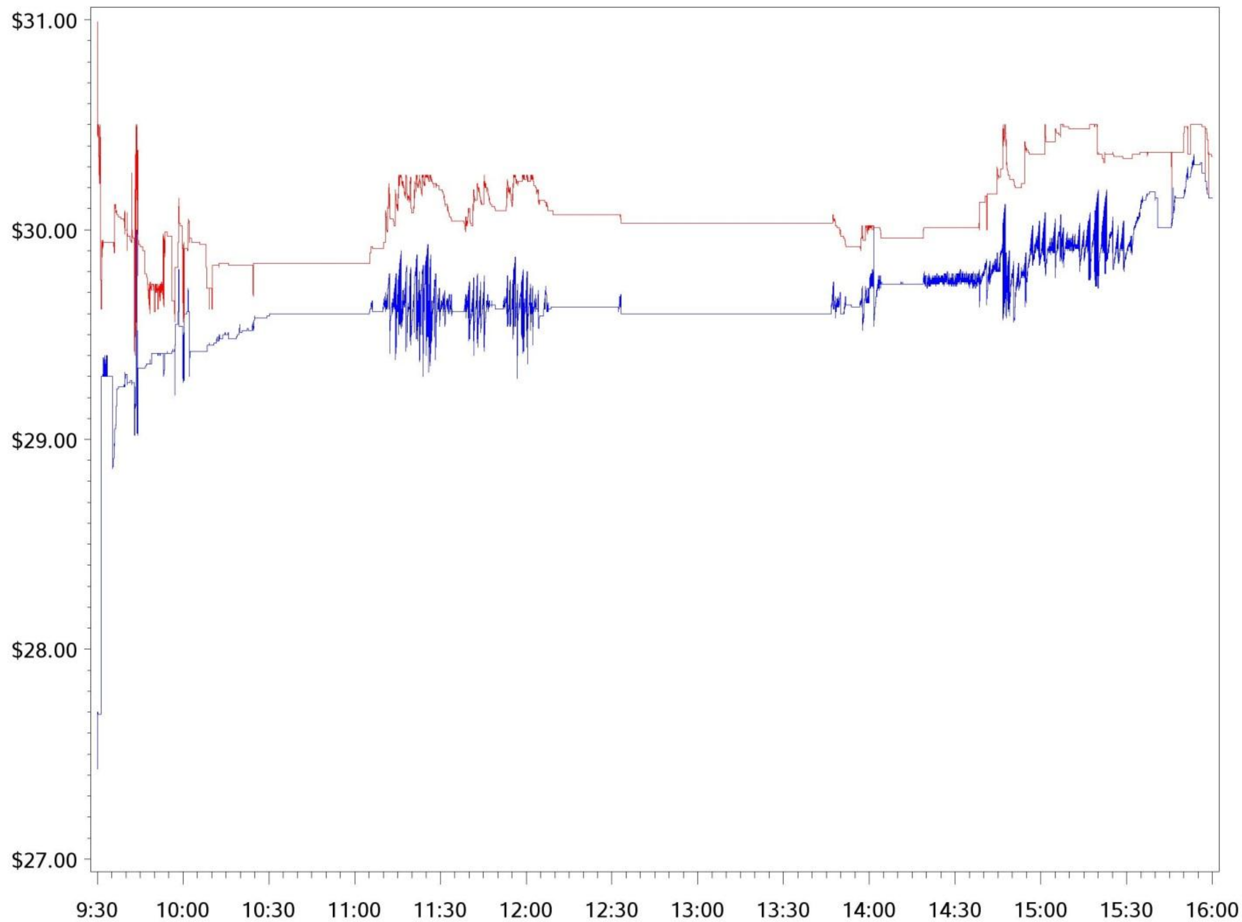

Figure 2. Wavelet variance ratios for the National Best Bid and Offer

The results are based on a random sample of 100 US stocks, stratified by dollar trading volume, for April, 2011. Wavelet variance estimates $\nu^2(\tau_j)$ by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer at one-millisecond resolution, and (for time scales of a second or longer) at a one-second resolution. The estimates are averaged across the NBB and NBO. The wavelet variance ratio is

$$VR(\tau_j) = \frac{\nu^2(\tau_j)}{\nu^2(\tau_J)} \times \frac{\nu^2_{\overline{rw}}(\tau_J)}{\nu^2_{rw}(\tau_j)}$$

$\nu^2(\tau_J)$ is the wavelet variance at the longest time scale in the analysis (34.1 minutes). The normalization ratio $\nu^2_{\overline{rw}}(\tau_J)/\nu^2_{rw}(\tau_j)$ is based on the assumption that the prices follow random-walks at a one millisecond resolution. The "$\overline{rw}$" subscript indicates that the variance has also been adjusted for averaging at the one-second resolution. Due to the normalization, $VR(\tau_j = 34.1\ min)$ is unity by construction. Plotted values are cross-firm medians for each of the dollar volume quintiles. Circles mark values in which the base resolution of the transforms was one millisecond; triangles indicate values with base resolution of one second.

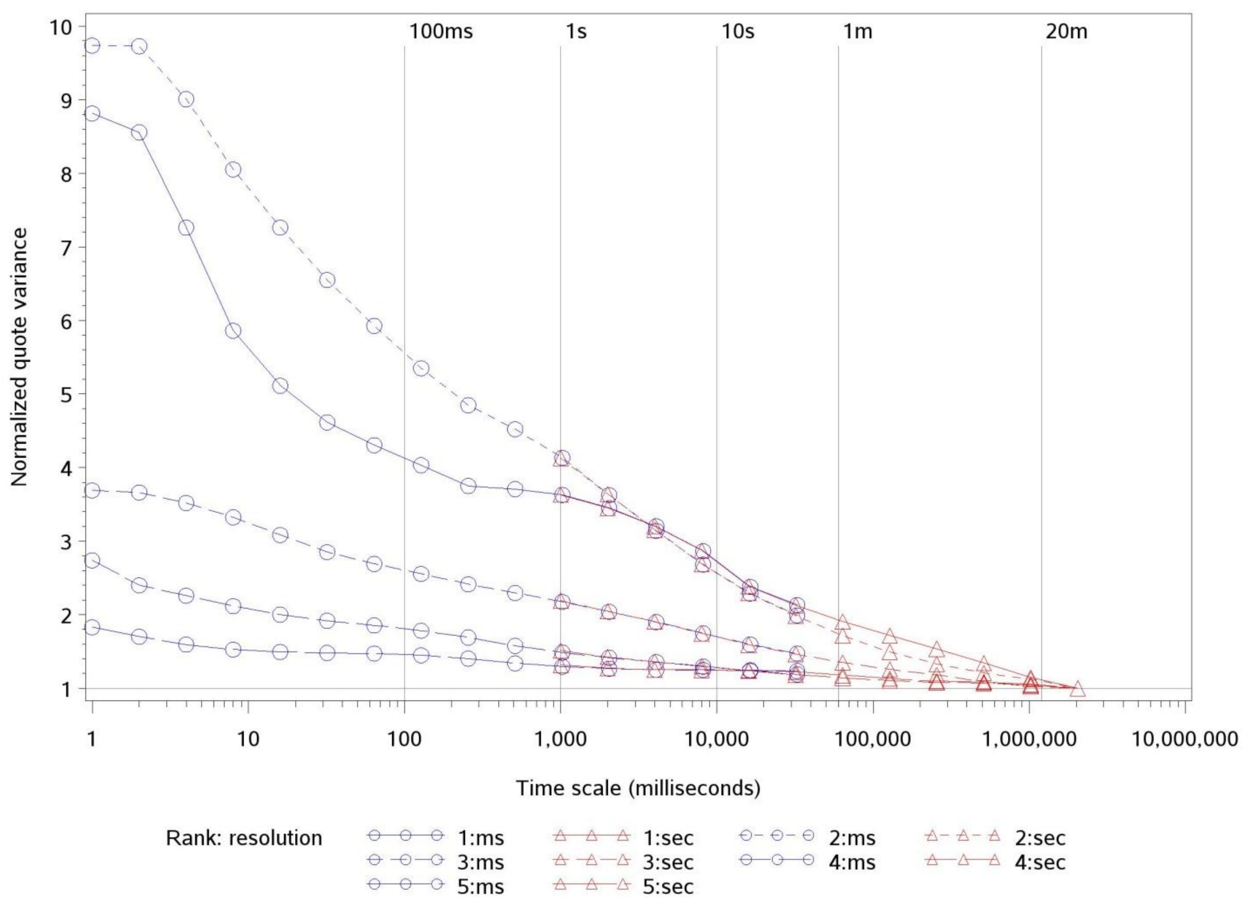

Figure 3. Wavelet correlations between the National Best Bid and National Best Offer

The results are based on a random sample of 100 US stocks, stratified by dollar trading volume, for April, 2011. Wavelet variance and covariance estimates by firm and time scale are formed using Haar MODWTs applied separately to the National Best Bid and the National Best Offer at one-millisecond resolution, and (for time scales of a second or longer) at a one-second resolution. Denoting these wavelet variances and covariances by $\nu_{NBB}^2(\tau_j)$, $\nu_{NBO}^2(\tau_j)$, $\nu_{NBB,NBO}(\tau_j)$, the wavelet correlated is defined as $\nu_{NBB,NBO}(\tau_j)/\sqrt{\nu_{NBB}^2(\tau_j) \times \nu_{NBO}^2(\tau_j)}$. Plotted values are cross-firm medians for each of the dollar volume quintiles. Circles mark values in which the base resolution of the transforms was one millisecond; triangles indicate values with base resolution of one second.