Does Anything Beat 5-Minute RV?

A Comparison of Realized Measures Across Multiple Asset Classes*

Lily Liu[†], Andrew J. Patton[†] and Kevin Sheppard[‡]

[†]Duke University and [‡]University of Oxford

December 5, 2012

Abstract

We study the accuracy of a wide variety of estimators of asset price variation constructed from high-frequency data (so-called "realized measures"), and compare them with a simple "realized variance" (RV) estimator. In total, we consider almost 400 different estimators, applied to 11 years of data on 31 different financial assets spanning five asset classes, including equities, equity indices, exchange rates and interest rates. We apply data-based ranking methods to the realized measures and to forecasts based on these measures. When 5-minute RV is taken as the benchmark realized measure, we find little evidence that it is outperformed by any of the other measures. When using inference methods that do not require specifying a benchmark, we find some evidence that more sophisticated realized measures significantly outperform 5-minute RV. In forecasting applications, we find that a low frequency "truncated" RV outperforms most other realized measures. Overall, we conclude that it is difficult to significantly beat 5-minute RV.

Keywords: realized variance, volatility forecasting, high frequency data.

J.E.L. classifications: C58, C22, C53.

*We thank Yacine Aït-Sahalia, Tim Bollerslev, Jia Li, Asger Lunde, George Tauchen, Julian Williams, and seminar participants at Cass Business School, Duke University, the conference in honor of Timo Teräsvirta at Aarhus University and the Ultra High Frequency Econometrics workshop on the Isle of Skye for helpful comments. Contact address: Andrew Patton, Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham NC 27708-0097. Email: andrew.patton@duke.edu.

1 Introduction

In the past fifteen years, many new estimators of asset return volatility constructed using high frequency price data have been developed (see Andersen et al. (2006), Barndorff-Nielsen and Shephard (2007) and Meddahi et al. (2011), *inter alia*, for recent surveys and collections of articles). These estimators generally aim to estimate the quadratic variation or the integrated variance of a price process over some interval of time, such as one day or week. We refer to estimators of this type collectively as "realized measures." This area of research has provided practitioners with an abundance of alternatives, inducing demand for some guidance on which estimators to use in empirical applications. In addition to selecting a particular estimator, these nonparametric measures often require additional choices for their implementation. For example, the practitioner must choose the sampling frequency to use and whether to sample prices in calendar time (every x seconds) or tick time (every x trades). When both transaction and quotation prices are available, the choice of which price to use also arises. Finally, some realized measures further require choices about tuning parameters such as a kernel bandwidth or "block size."

The aim of this paper is to provide guidance on the choice of realized measure to use in applications. We do so by studying the performance of a large number of realized measures across a broad range of financial assets. In total we consider almost 400 realized measures, across seven distinct classes of estimators, and we apply these to 11 years of daily data on 31 individual financial assets covering five asset classes. We compare the realized measures in terms of their estimation accuracy for the latent true quadratic variation, and in terms of their forecast accuracy when combined with a simple and well-known forecasting model. We employ model-free data-based comparison methods that make minimal assumptions on properties of the efficient price process or on the market microstructure noise that contaminates the efficient prices.

The fact that the target variable in this analysis (quadratic variation) is latent, even ex-post, creates an obstacle to applying standard techniques. Previous theoretical research on the selection of estimators of quadratic variation has often focused on recommending a sampling frequency, or other tuning parameter, based on the underlying theory using plug-in type estimators of nuisance parameters. For some estimators, a formula for the optimal sampling frequency under a set of assumptions is derived and can be computed using estimates of higher order moments, see Bandi and Russell (2008) among others. However, these formulas are usually heavily dependent on assumptions about the microstructure noise and efficient price process, such as independence of the noise from the price, serial correlation, etc.

Previous empirical work on the choice of realized measure has been based on a relatively homogeneous collection of assets (most commonly, constituents of the Dow Jones Industrial Average index) or on results from simulations, see Aït-Sahalia and Mancini (2008), Gatheral and Oomen (2010), Andersen et al. (2011) and Ghysels and Sinko (2011). The benefit of using simulations is that the true volatility is known to the researcher, and no inference is required to rank the alternative realized measures; the drawback is the potential sensitivity of the results to specific choices for the price process or the noise process. Our analysis extends previous work on this topic by considering a large, relatively heterogeneous collection of assets, which provides an opportunity to compare realized measures in environments with varied price processes and market microstructures. By using real and varied financial data, we avoid having to specify any form for the market microstructure process, which could lead to a bias in favor of one or another realized measure.

Our objective is to compare a large number of available realized measures in a unified, data-based, framework. We use the data-based ranking method of Patton (2011a), which makes no assumptions about the properties of the market microstructure noise, aside from standard moment and mixing conditions. The main contribution of this paper is an empirical study of the relative performance of estimators of daily quadratic variation from 7 types of realized measures using data from 31 financial assets spanning different classes. We use transactions and quotations prices from January 2000 to December 2010, sampled in calendar time and tick-time, for many sampling frequencies ranging from 1 second to 15 minutes. We use the "model confidence set"of Hansen et al. (2011) to construct sets of realized measures that contain the best measure with a given level of confidence. We are also interested whether a simple RV estimator with a reasonable choice of sampling frequency, namely 5-minute RV, can stand in as a "good enough" estimator for QV. This is similar to the comparison of more sophisticated volatility models with a simple benchmark model presented in Hansen and Lunde (2005). We use the step-wise multiple testing method of Romano and Wolf (2005), which allows us to determine whether any of the 390 or so competing realized measures is significantly more accurate than a simple realized variance measure based on 5-minute returns. We also conduct an out-of-sample forecasting experiment to study the accuracy of volatility forecasts based on these individual realized measures, when used in the "heterogeneous autoregressive" (HAR) forecasting model of Corsi (2009), for forecast horizons ranging from 1 to 50 trading days.

The remainder of this paper is organized as follows. Section 2 provides a brief description of the classes of realized measures. Section 3 describes ranking methodology and tests used to compare the realized measures. Section 4 describes the high frequency data and the set of realized measures we construct. Our main analysis is presented in Section 5, and Section 6 concludes.

2 Measures of asset price variability

To fix ideas and notation, consider a general jump-diffusion model for the log-price p of an asset:

$$dp\left(t\right) = \mu\left(t\right)dt + \sigma\left(t\right)dW\left(t\right) + \kappa\left(t\right)dN\left(t\right) \tag{1}$$

where μ is the instantaneous drift, σ is the (stochastic) volatility, W is a standard Brownian motion, κ is the jump size, and N is a counting measure for the jumps. In the absence of jumps the third term on the right-hand side above is zero. The quadratic variation of the log-price process over period $t+1$ is defined

$$\begin{aligned} QV_{t+1} &= \operatorname*{plim}_{n\to\infty}\sum_{j=1}^{n} r_{t+j/n}^{2} \\ \text{where}\quad r_{t+j/n} &= p_{t+j/n} - p_{t+(j-1)/n} \end{aligned} \tag{2}$$

where the price series on day $t+1$ is assumed to be observed on a grid of n times $\left\{p_{t+1/n},...,p_{t+1-1/n},p_{t+1}\right\}$. See Andersen et al. (2006) and Barndorff-Nielsen and Shephard (2007) for surveys of volatility estimation and forecasting using high frequency data. The objective of this paper is to compare the variety of estimators of QV that have been proposed in the literature to date. We do so with emphasis on comparisons with the simple realized variance estimator, which is the empirical analog of QV:

$$RV_{t+1} = \sum_{j=1}^{n} r_{t+j/n}^2 .$$

2.1 Sampling frequency, sampling scheme, and sub-sampling

We consider a variety of classes of estimators of asset price variability. All realized measures require a choice of sampling frequency (e.g., 1-second or 5-minute sampling), sampling scheme (calendar time or tick time), and, for most assets, whether to use transaction prices or mid-quotes. Thus even for a very simple estimator such as realized variance, there are a number of choices to be made. To examine the sensitivity of realized measures to these choices, we implement each measure using calendar-time sampling of 1 second, 5 seconds, 1 minute, 5 minutes and 15 minutes. We also consider tick-time sampling using samples that yield *average* durations that match the values for calendar-time sampling, as well as a "tick-by-tick" estimator that simply uses every available observation. Subsampling,[1] introduced by Zhang et al. (2005), is a simple way to improve efficiency of some sparse-sampled estimators. We consider subsampled versions of all the estimators (except estimators using tick-by-tick data, which cannot be subsampled).[2] The sub-sampled version of RV (which turns out to perform very well in our analysis), was first studied as the "second best" estimator in Zhang et al. (2005), and is called the "average RV" estimator in Andersen et al. (2011) and Ghysels and Sinko (2011).

In total we have 5 calendar-time implementations, 6 tick-time implementations, and 5+6–1=10 corresponding subsampled implementations, yielding 21 realized measures for a given price series. Estimating these on both transaction and quote prices yields a total of 42 versions of each realized measure. Of course, some of these combinations are expected to perform poorly empirically (given the extant literature on microstructure biases and the design of some of the estimators described

[1]Subsampling involves using a variety of "grids" of prices sampled at a given frequency to obtain a collection of realized measures, which are then averaged to yield the "subsampled" version of the estimator. For example, 5-minute RV can be computed using prices sampled at 9:30, 9:35, etc. and can also be computed using prices sampled at 9:31, 9:36, etc.

[2]In general, we implement subsampling using a maximum of 10 partitions.

below), and by including them in our analysis we thus have an "insanity check" on whether our tests can identify these poor estimators.

2.2 Classes of realized measures

The first class of estimators is standard realized variance (RV), which is the sum of squared intra-daily returns. This simple estimator is the sample analog of quadratic variation, and in the absence of noisy data, it is the nonparametric maximum likelihood estimator, and so is efficient, see Andersen et al. (2001b) and Barndorff-Nielsen (2002). However, market microstructure noise induces serial auto-correlation in the observed returns, which biases the realized variance estimate at high sampling frequencies (see Hansen and Lunde (2006b) for a detailed analysis of the effects of microstructure noise). When RV is implemented in practice, the price process is often sampled sparsely to strike a balance between increased accuracy from using higher frequency data and the adverse effects of microstructure noise. Popular choices include 1-minute, 5-minute (as in the title of this paper), or 30-minute sampling.

We next draw on the work of Bandi and Russell (2008), who propose a method for optimally choosing the sampling frequency to use with a standard RV estimator. This sampling frequency is calculated using estimates of integrated quarticity[3] and variance of the microstructure noise. These authors also propose a bias-corrected estimator that removes the estimated impact of market microstructure noise. Since the key characteristic of the Bandi-Russell estimator is the estimated optimal sampling frequency, we do not vary the sampling frequency when implementing it. This reduces the number of versions of this estimator from 42 to 8.

The third class of realized measures we consider is the first-order autocorrelation-adjusted RV estimator (RVac1) used by French et al. (1987) and Zhou (1996), and studied extensively by Hansen and Lunde (2006b). This estimator was designed to capture the effect of autocorrelation in high frequency returns induced by market microstructure noise.

The fourth class of realized measures includes the two-scale realized variance (TSRV) of Zhang et al. (2005) and the multi-scale realized variance (MSRV) of Zhang (2006). These estimators

[3]Estimates of daily integrated quarticity are estimated using 39 intra-day prices sampled uniformly in tick-time.

compute a subsampled RV on one or more slower time scales (lower frequencies) and then combine with RV calculated on a faster time scale (higher frequency) to correct for microstructure noise. Under certain conditions on the market microstructure noise, these estimators are consistent at the optimal rate. In our analysis, we set the faster time scale by using one of the 21 sampling frequency/sampling scheme combinations mentioned above, while the slower time scale(s) are chosen to minimize the asymptotic variance of the estimatorm using the methods developed in the original papers. It is worth noting here that "subsampled RV", which we have listed in our first class of estimators, corresponds to the "second-best" form of TSRV in Zhang et al. (2005), in that it exploits the gains from subsampling but does not attempt to estimate and remove any bias in this measure. We keep any measure involving two or more time scales in the TSRV/MSRV class, and any measures based on a single time scale are listed in the RV class.

The fifth class of realized measures is the realized kernel (RK) estimator of Barndorff-Nielsen et al. (2008). This measure is a generalization of RVac1, accommodating a wider variety of microstructure effects and leading to a consistent estimator. Barndorff-Nielsen et al. (2008) present realized measures using several different kernels, and we consider RK with the "flat top" versions of the Bartlett, cubic, and modified Tukey-Hanning$_2$ kernel, and the "non-flat-top"Parzen kernel. The Bartlett and cubic kernels are asymptotically equivalent to TSRV and MSRV, and modified Tukey-Hanning$_2$ was the recommended kernel in Barndorff-Nielsen et al. (2008) in their empirical application to GE stock returns. The non-flat-top Parzen kernel was studied further in Barndorff-Nielsen et al. (2011) and results in a QV-estimator that is always positive while allowing for dependence and endogeneity in the microstructure noise. We implement these realized kernel estimators using the 21 sampling frequency/sampling scheme combinations mentioned above, and estimate the optimal bandwidths for these kernels separately for each day, using the methods in Barndorff-Nielsen et al. (2011). The realized kernel estimators are not subsampled because Barndorff-Nielsen et al. (2011) report that for "kinked" kernels such as the Bartlett kernel, the effects of subsampling are neutral, while for the other three "smooth" kernels, subsampling is detrimental. (The RVac1 measure corresponds to the use of a "truncated" kernel, and subsampling improves performance, so we include the subsampled versions of RVac1 in the study.)

The sixth class of estimators is the "realized range-based variance" (RRV) of Christensen and Podolskij (2007) and Martens and Van Dijk (2007). Early research by Parkinson (1980), Andersen and Bollerslev (1998) and Alizadeh et al. (2002) show that the properly scaled, daily high-low range of log prices is an unbiased estimator of daily volatility when constant, but is more efficient than squared daily open-to-close returns. Correspondingly, Christensen and Podolskij (2007) and Martens and Van Dijk (2007) apply the same arguments to intra-day data, and improve on the RV estimator by replacing each intra-day squared return with the high-low range from a block of intra-day returns. To implement RRV, we use the sampling schemes described above, and then use block size of 5, following Patton and Sheppard (2009a), and block size of 10, which is close to the average block size used in Christensen and Podolskij's application to General Motors stock returns.

Finally, we include the maximum likelihood Realized Variance (MLRV) of Aït-Sahalia et al. (2005), which assumes that the observed price process is composed of the efficient price plus i.i.d. noise such that the observed return process follows an MA(1) process, with parameters that can be estimated using Gaussian MLE. This estimator is shown to be robust to misspecification of the marginal distribution of the microstructure noise by Aït-Sahalia et al. (2005), but is sensitive to the independence assumption of noise, as demonstrated in Gatheral and Oomen (2010).

The total number of realized measures we compute for a single price series is 199, so an asset with both transactions and quote data has a set of 398 realized measures.[4]

2.3 Additional realized measures

Our main empirical analysis focuses on realized measures that estimate the quadratic variation of an asset price process. From a forecasting perspective, work by Andersen et al. (2007) and others has shown that there may be gains to decomposing QV into the component due to continuous

[4]Specifically, for RV, TSRV, MSRV, MLRV, RVac1, RRV (with two choices of block size) and RK (with 4 different kernels), 11 not-subsampled estimators, which span different sampling frequencies and sampling schemes, are implemented on each of the transactions and midquotes price series. In addition, we estimate 2 bias-corrected Bandi-Russell realized measures and 2 not-bias-corrected BR measures (calendar-time and tick-time sampling) per price series. These estimators account for $11\times11\times2 + (2+2)\times2 = 250$ of the total set. RV, TSRV, MSRV, MLRV, RVac1 and RRV (m=5 and 10) also have 10 subsampled estimators per price series, and there are 4 subsampled BR estimators per price series, which adds $7\times10\times2 + 4\times2 = 148$ subsampled estimators to the set. In total, this makes 250+148=398 estimators.

variation (integrated variance, or IV) and the component due to jumps (denoted JV):

$$QV_{t+1} = \operatorname*{plim}_{n\to\infty} \sum_{j=1}^{n} r_{t+j/n}^2 = \underbrace{\int_t^{t+1} \sigma^2(s)\, ds}_{IV_{t+1}} + \underbrace{\sum_{t<s\leq t+1} \kappa^2(s)}_{JV_{t+1}} \tag{3}$$

Thus for our forecasting application in Section 5.6, we also consider four classes of realized measures that are "jump robust", i.e., they estimate IV not QV. The first of these is the bi-power variation (BPV) of Barndorff-Nielsen and Shephard (2006), which is a scaled sum of products of adjacent absolute returns. The second class of jump-robust realized measures is the quantile-based realized variance (QRV) of Christensen et al. (2010). The QRV is based on combinations of locally extreme quantile observations within blocks of intra-day returns, and requires choice of block length and quantiles. It reported to have better finite sample performance than BPV in the presence of jumps, and additionally is consistent, efficient and jump-robust even in the presence of microstructure noise. For implementation, we use the asymmetric version of QRV with rolling overlapping blocks[5] and quantiles approximately equal to 0.85, 0.90 and 0.96, following their empirical application to Apple stock returns. The block lengths are chosen to be around 100, with the exact value depending on the number of filtered daily returns, and the quantile weights are calculated optimally following the method in Christensen et al. (2010). QRV is the most time-consuming realized measure to estimate, and thus is not further subsampled.

The third class of jump-robust realized measures are the "nearest neighbor truncation" estimators of Andersen et al. (2008), specifically their "MinRV" and "MedRV" estimators. These are the scaled square of the minimum of two consecutive intra-day absolute returns or the median of 3 consecutive intra-day absolute returns. These estimators are more robust to jumps and microstructure noise than BPV, and MedRV is designed to handle outliers or incorrectly entered price data.

The final class of jump-robust measures estimators is the truncated or threshold realized variance (TRV) of Mancini (2009, 2001), which is the sum of squared returns, but only including returns

[5]Christensen et al. (2010) refers to this formulation of the QRV as "subsampled QRV", as opposed to "block QRV", which has adjacent non-overlapping blocks. However, we do not use this terminology as this type of "subsampling" is different from the subsampling we implement for the other estimators.

that are smaller in magnitude than a certain threshold. We take the threshold to be $4\sqrt{n^{-1}BPV_t}$, where n is the number of sampled intra-day returns and BPV_t is the previous day's bi-power estimate using 1-minute calendar-time sampling of transactions prices.

In total, across sampling frequencies and subsampling/not subsampling we include 206 jump-robust realized measures in our forecasting application, in addition to the 398 estimators described in the previous section.

3 Comparing the accuracy of realized measures

We examine the empirical accuracy of our set of competing measures of asset price variability using two complementary approaches.

3.1 Comparing estimation accuracy

We first compare the accuracy of realized measures in terms of their estimation error for a given day's quadratic variation. QV is not observable, even *ex post*, and so we cannot directly calculate a metric like mean-squared error and use that for the comparison. We overcome this by using the data-based ranking method of Patton (2011a). This approach requires employing a proxy (denoted $\tilde{\theta}$) for the quadratic variation that is assumed to be unbiased, but may be noisy.[6] This means that we must choose a realized measure that is unlikely to be affected by market microstructure noise. Using proxies that are more noisy will reduce the ability to discriminate between estimators, but will not affect consistency of the procedure. We use the squared open-to-close returns from trades prices (RVdaily) for our main analysis, and further consider 15-minute RV, 5-minute RV, 1-minute MSRV and 1-minute RKth2, all computed on trades prices using tick-time sampling, as possible alternatives.[7,8] Since estimators based on the same price data are correlated, it is necessary to use a

[6] Numerous estimators of quadratic variation can be shown to be asymptotically unbiased, as the sampling interval goes to zero, however this approach requires unbiasedness for a fixed sampling interval.

[7] These four other proxies were found to be unbiased for the RVdaily measure for the majority of assets, and in addition, are generally much more precise.

[8] We use volatility proxies from different classes of realized measures (RV, MSRV and RK) to reassure the reader that the rankings we obtain are not sensitive to the choice of proxy. Subject to the proxy being unbiased, the choice of proxy should not (asymptotically) affect the rankings and this is indeed confirmed in our empirical results in Section 5.

lead (or a lag) of the proxy to "break" the dependence between the estimation error in the realized measure under analysis and the estimation error in the proxy. We use a one-day lead.[9]

The comparison of estimation accuracy also, of course, requires a metric for measuring accuracy. The approach of Patton (2011a) allows for a variety of metrics, including the MSE and QLIKE loss functions. Simulation results in Patton and Sheppard (2009b), and empirical results in Hansen and Lunde (2005), Patton and Sheppard (2009a) and Patton (2011a) all suggest that using QLIKE leads to more power to reject inferior estimators. The QLIKE loss function is defined

$$\text{QLIKE} \qquad L(\theta, M) = \frac{\theta}{M} - \log\frac{\theta}{M} - 1 \tag{4}$$

where θ is QV, or a proxy for it, and M is a realized measure. With this in hand, we obtain a consistent (as $T \to \infty$) estimate of the difference in accuracy between any two realized measures:

$$\frac{1}{T}\sum_{t=1}^{T} \Delta\tilde{L}_{ij,t} \xrightarrow{p} E\left[\Delta L_{ij,t}\right] \tag{5}$$

where $\Delta\tilde{L}_{ij,t} \equiv L\left(\tilde{\theta}_t, M_{it}\right) - L\left(\tilde{\theta}_t, M_{jt}\right)$ and $\Delta L_{ij,t} \equiv L(\theta_t, M_{it}) - L(\theta_t, M_{jt})$. Under standard regularity conditions (see Patton (2011a) for example) we can use a block bootstrap to conduct tests on the estimated differences in accuracy, such as the pair-wise comparisons of Diebold and Mariano (2002) and Giacomini and White (2006), the "reality check" of White (2000) as well as the multiple testing procedure of Romano and Wolf (2005), and the "model confidence set" of Hansen et al. (2011).

[9] As described in Patton (2011a), the use of a lead (or lag) of the proxy formally relies on the daily quadratic variation following a random walk. Numerous papers, see Bollerslev et al. (1994) and Andersen et al. (2006) for example, find that conditional variance is a very persistent process, close to being a random walk. Hansen and Lunde (2010) study the quadratic variation of all 30 constituents of the Dow Jones Industrial Average and reject the null of a unit root for almost none of the stocks. Simulation results in Patton (2011a) show that inference based on this approach has acceptable finite-sample properties for DGPs that are persistent but strictly not random walks, and we confirm in Table A4, in the appendix, that all series studied here are highly persistent.

3.2 Comparing forecast accuracy

The second approach we consider for comparing realized measures is through a simple forecasting model. As we describe in Section 5.6 below, we construct volatility forecasts based on the heterogeneous autoregressive (HAR) model of Corsi (2009), estimated separately for each realized measure. The problem of evaluating volatility forecasts has been studied extensively, see Hansen and Lunde (2005), Andersen et al. (2005), Hansen and Lunde (2006a) and Patton (2011b) among several others. The latter two papers focus on applications where an unbiased volatility proxy is available, and again under standard regularity conditions we can use block bootstrap methods to conduct tests such as those of Diebold and Mariano (2002), White (2000), Romano and Wolf (2005), Giacomini and White (2006), and Hansen et al. (2011).

4 Data description

We use high frequency (intra-daily) asset price data for 31 assets spanning five asset classes: individual equities (from the U.S. and the U.K.), equity index futures, computed stock indices, currency futures and interest rate futures. The data are transactions prices and quotations prices taken from Thomson Reuter's Tick History. The sample period is January 2000 to December 2010, though data availability limits us to a shorter sub-period for some assets. Short days, defined as days with prices recorded for less than 60% of the regular market operation hours, are omitted. For each asset, the number of short days is small compared to the total number of days – the largest proportion of days omitted is 1.7% for ES (E-mini S&P500 futures). Across assets, we have an average of 2537 trading days, with the shortest sample being 1759 trade days (around 7 years) and the longest 2782 trade days. All series were cleaned according to a set of baseline rules similar to those in Barndorff-Nielsen et al. (2009). Data cleaning details are provided in the appendix.

Table 1 presents the list of assets, along with their sample periods and some summary statistics. Computed stock indices are not traded assets and are constructed using trade prices, and so quotes are unavailable. This table reveals that these assets span not only a range of asset classes, but also characteristics: average annualized volatility ranges from under 2%, for interest rate futures, to over 40%, for individual equities. The average time between price observations ranges from under one second, for the E-mini S&P 500 index futures contract, to nearly one minute, for some individual equities and computed equity indices.

[INSERT TABLE 1 ABOUT HERE]

Given the large number of realized measures and assets, it is not feasible to present summary statistics for all possible combinations. Table A1 in the appendix describes the shorthand used to describe the various estimators[10], and in Table 2 we present summary statistics for a selection of realized measures for two assets, Microsoft and the US dollar/Australian dollar futures contract.[11] Tables A3 and A4 in the appendix contain more detailed summary statistics. Table 2 reveals some familiar features of realized measures: those based on daily squared returns have similar averages to realized measures using high (but not too high) frequency data, but are more variable, reflecting greater measurement error. For Microsoft, for example, RVdaily has an average of 3.20 (28.4% annualized) compared with 3.37 for RV5min, but its standard deviation is more than 25% larger than that of RV5min. We also note that RV computed using tick-by-tick sampling (i.e., the highest possible sampling) is much larger on average than the other estimators, more than 3 times larger for Microsoft and around 50% larger for the USD/AUD exchange rate, consistent with the presence of market microstructure noise.

In the last four columns of Table 2 we report the first- and second-order sample autocorrelations of the realized measures, as well as estimates of the first- and second-order autocorrelation of the underlying quadratic variation using the estimation method in Hansen and Lunde (2010).[12] As expected, the latter estimates are much higher than the former, reflecting the attenuation bias due to the estimation error in a realized measure. Using the method of Hansen and Lunde (2010), the estimated first-order autocorrelation of QV for Microsoft and the USD/AUD exchange rate is around 0.95, while the sample autocorrelation for the realized measures themselves averages

[10] For example, "RV_1m_ct_ss" refers to realized variance (RV), computed on 1-minute data (1m) sampled in calendar time (c), using trade prices (t), with subsampling (ss). See Table A1 for details.

[11] All realized measures were computed using code based on Kevin Sheppard's "Oxford Realized" toolbox for Matlab, http://realized.oxford-man.ox.ac.uk/data/code.

[12] Following their empirical application to the 30 DJIA stocks, we use the demeaned 4th through 10th lags of the daily QV estimator as instruments.

around 0.68. Table A4 presents summaries of these autocorrelations for all 31 assets, and reveals that the estimated first- (second-) order autocorrelation of the underlying QV is high for all assets. The average estimate across assets realized measures, even the poor estimators, equals 0.95 (0.92). These findings support our use, in the next section, of the ranking method of Patton (2011a), which relies on high persistence of QV.

[INSERT TABLE 2 ABOUT HERE]

5 Empirical results on the accuracy of realized measures

We now present the main analysis of this paper. We firstly discuss simple rankings of the realized measures, and then move on to more sophisticated tests to formally compare the various measures. As described in Section 3, we measure accuracy using the QLIKE distance measure, using squared open-to-close returns (RVdaily) as the volatility proxy, with a one-day lead to break the dependence between estimation error in the realized measure and error in the proxy. In some of the analysis below we consider using higher frequency RV measures for the proxy (RV15min and RV5min), as well as some non-RV proxies, namely 1-minute MSRV and 1-minute Tukey-Hanning$_2$ Realized Kernel.

5.1 Rankings of average accuracy

We firstly present a summary of the rankings of the 398 realized measures applied to the 31 assets in our sample. These rankings are based on average, unconditional distance of the measure from the true QV, and in Section 5.5 we consider conditional rankings.

The top panel of Table 3 presents the "top 10" individual realized measures, according to their average rank across all assets in a given class.[13] It is noteworthy that 5-minute RV does *not* appear in the top 10 for any of these asset classes. This is some initial evidence that there are indeed better

[13]Table A6 in the appendix presents rank correlation matrices for each asset class, and confirms that the rankings of realized measures for individual assets in a given asset class are relatively consistent, with rank correlations ranging from 0.68 to 0.87.

estimators of QV available, and we test whether this outperformance is statistically significant in the sections below.

With the caveat that these estimated rankings do not come with any measures of significance, and that realized measures in the same class are likely highly correlated[14], we note the following patterns in the results. Realized kernels appear to do well for individual equities (taking 7 of the top 10 slots), realized range does well for interest rate futures (8 out of top 10), and two/multi-scales RV do well for currency futures (6 out of the top 10). For computed indices, RVac1 and realized kernels comprise the entire top 10. The top 10 realized measures for index futures contain a smattering of measures across almost all classes. The lower panel of Table 3 presents a summary of the upper panel, sorting realized measures by class and sampling frequency.

It is perhaps also interesting to note which price series is most often selected. We observe a mix of trades and quotes for individual equities,[15] while we see mid-quotes dominating the top 10 for interest rate futures and currency futures. For equity index futures, transaction prices make up the entire top 10. (Our computed indices are only available with transaction prices, so no comparisons are available for that asset class.)

[INSERT TABLE 3 ABOUT HERE]

5.2 Pair-wise comparisons of realized measures

To better understand the characteristics of a "good" realized measure, we present results on pair-wise comparisons of measures that differ only in one aspect. We consider three features: the use of calendar-time vs. tick-time sampling; the use of transaction prices vs. mid-quotes; and the use of subsampling. For each class of realized measures and for each sampling frequency, we compare pairs of estimators that differ in one of these dimensions, and compute a robust t-statistic on the average difference in loss, separately for each asset.[16] Table 4 presents the proportion (across the

[14]See Table A5 in the appendix for a summary of the correlations between realized measures.

[15]In fact, decomposing this group into US equities and UK equities, we see that the top 10 realized measures for US equities all use transaction prices, while the top 10 for UK equities all use mid-quotes, perhaps caused by different forms of market microstructure noise on the NYSE and the LSE.

[16]This is done as a panel regression for a single asset, as for each measure of a specific estimator class and sampling frequency, there are $2 \times 2 \times 2 = 8$ versions (cal-time vs. tick time, trades vs. quotes, not subsampled vs. subsampled), and conditioning on one of these characteristics leaves 4 versions.

31 assets) of t-statistics that are significantly positive minus the proportion that are significantly negative.[17] A negative entry in a given element indicates that the first approach (eg, calendar-time sampling in the top panel) outperforms the second approach.

The top panel of Table 4 reveals that for high frequencies (1-second and 5-second), calendar time sampling is preferred to tick-time sampling, while for lower frequencies (5-minute and 15-minute), tick-time sampling generally leads to better realized measures. Oomen (2006) and Hansen and Lunde (2006c) provide theoretical grounds for why tick-time sampling should outperform calendar-time sampling, and at lower frequencies this appears to be true. Microstructure noise may (likely) play a role at the highest frequencies, and the ranking of calendar-time and tick-time sampling depends on their sensitivity to this noise.

The middle panel of Table 4 shows that transaction prices are generally preferred to quote prices for most estimator-frequency combinations. Exceptions are RV, MLRV and RRV at the highest frequencies (1-tick and/or 1-second) and MSRV at low frequencies.

The lower panel of Table 4 compares realized measures with and without subsampling. Theoretical work by Zhang et al. (2005), Zhang (2006), Andersen et al. (2011) and Ghysels and Sinko (2011) suggests that subsampling is a simple way to improve the efficiency of a realized measure. Our empirical results generally confirm that subsampling is helpful, at least when using lower frequency (5-minute and 15-minute) data. For higher frequencies (1-second to 1-minute), subsampling has either a neutral or negative impact on accuracy. Interestingly, we note that for the realized range (RRV), subsampling reduces accuracy across all sampling frequencies.

[INSERT TABLE 4 ABOUT HERE]

5.3 Does anything beat 5-minute RV?

Realized variance, computed with a reasonable choice of sampling frequency, is often taken as a benchmark or rule-of-thumb estimator for volatility, see Andersen et al. (2001a) and Barndorff-

[17] The format of the panels in this table vary slightly: the top panel does not have a column for 1-tick sampling as there is no calendar-time equivalent, and the lower panel does not have this column as 1-tick measures cannot be subsampled. The lower panel does not contain the RK row, given the work of Barndorff-Nielsen et al. (2011). Finally, the middle panel covers only 26 assets, as we only have transaction prices for the 5 computed indices.

Nielsen (2002) for example. This measure has been used as far back as French et al. (1987), is simple to compute, and when implemented on a relatively low sampling frequency (such as 5-minutes) requires much less data and data cleaning. Thus it is of great interest to know whether it is significantly outperformed by one of the many more sophisticated realized measures proposed in the literature.

We use the stepwise multiple testing method of Romano and Wolf (2005) to address this question. The Romano-Wolf method tests the unconditional accuracy of a set of estimators relative to that of a benchmark realized measure, which we take to be RV computed using 5-minute calendar time sampling on transaction prices (which we denote RV5min). This procedure is an extension of the "reality check" of White (2000), allowing us to determine not only *whether* the benchmark measure is rejected, but to identify the competing measures that led to the rejection. Formally, the Romano-Wolf stepwise method examines the set of null hypotheses:

$$H_0^{(s)} : E\left[L\left(\theta_t, M_{t,0}\right)\right] = E\left[L\left(\theta_t, M_{t,s}\right)\right], \text{ for } s = 1, 2, ..., S \tag{6}$$

and looks for realized measures, $M_{t,s}$, such that either $E\left[L\left(\theta_t, M_{t,0}\right)\right] > E\left[L\left(\theta_t, M_{t,s}\right)\right]$ or $E\left[L\left(\theta_t, M_{t,0}\right)\right] < E\left[L\left(\theta_t, M_{t,s}\right)\right]$. The Romano-Wolf procedure controls the "family-wise error rate", which is the probability of making one or more false rejections among the set of hypotheses. We run the Romano-Wolf test in both directions, firstly to identify the set of realized measures that are significantly worse than RV5min, and then to identify the set of realized measures that are significantly better than RV5min. We implement the Romano-Wolf procedure using the Politis and Romano (1994) stationary bootstrap with 1000 bootstrap replications, and an average block size of 10 days. A summary of results is presented in Table 5, and detailed results can be found in the online appendix.

The striking feature of Table 5 is the preponderance of estimators that are significantly beaten by RV5min, and the almost complete lack of estimators that significantly beat RV5min. Concerns about potential low power of this inference method are partially addressed by the ability of this method to reject so many estimators as significantly worse than RV5min: using daily RV as the proxy we reject an average of 185 estimators (out of 398) as significantly worse than RV5min, which represents approximately half of the set of competing measures.[18]

We also present results using the other four proxies. These proxies are more precise, although they are potentially more susceptible to market microstructure noise. Results from the more precise proxies are very similar: with these better proxies we can reject almost two-thirds of competing estimators as being significantly worse than RV5min, but we find just one asset out of 31 has *any* measures that significantly outperform RV5min.[19,20]

The asset for which we find that RV5min is significantly beaten, the 10-year US Treasury note futures contract (TY), is among the most frequently traded in our sample. (It is noteworthy, however, that there are five other assets that are more or comparably liquid but for which we find *no* realized measure significantly better than RV5min.[21]) For the 10-year Treasury note, the realized measures that outperform RV5min include MSRV, RK and RRV all estimated using 1-second or 5-second sampling (in calendar time or business time, with or without subsampling), and 1-minute RV and 1-minute RVac1; a collection of measures that one might expect to do well for a very liquid asset.

It is also noteworthy, that, combining the set of estimators that are significantly worse than RV5min (between a half and two-thirds of all estimators) with those that are significantly better

[18]We note here that averaging across all possible tuning parameters for a given estimator, as we do in Table 5, may obscure the good performance of an estimator using well-chosen tuning parameters, by grouping it with estimators using poorly-chosen tuning parameters. An alternative to this is pulling out a "reasonable" version of each estimator from the entire set, and considering only this reduced set of "reasonable" realized measures. The difficulty with this approach is determining *ex ante* the "reasonable" versions across assets with widely varying characteristics (e.g., computed stock indices vs. currency futures).

[19]We also implemented the Romano-Wolf procedure swapping the "reality check" step with a step based on the test of Hansen (2005). This latter test is designed to be less sensitive to poor alternatives with large variances (a potential concern in our application) and so should have better power. We found no change in the number of rejections. In a more forceful attempt to examine the sensitivity to poor alternatives: we identified, ex ante, 72 estimators that the existing literature would suggest are likely to have poor performance (for example, realized kernels on 15-minute returns). We removed this group of estimators from the competing set, and conducted the Romano-Wolf procedure on the remainder of the competing set. We found virtually no change in results of the tests – in fact, counting across the two Romano-Wolf tests for each of 31 assets, there was only one instance where an estimator was found to have different outcome from the original test.

[20]It is worth noting here that Table 5 reveals that the use of a particular measure does *not* lead to an apparent improvement in the performance of measures from the same class. Specifically, using a RV as the proxy does not "favor" RV measures, and using RK or TSRV does not favor RK or TSRV measures. The use of a one-day lead of the proxy solves this potential problem.

[21]These four assets are the futures contracts on S&P500, the FTSE 100, the EuroStoxx 50, the DAX 40 and the Euro/USD exchange rate.

(approximately zero), leaves between one-third and one-half of the set of 398 estimators that are not significantly different than RV5min in terms of average accuracy.

[INSERT TABLE 5 ABOUT HERE]

To better understand the results of the Romano-Wolf tests applied to this large collection of assets and realized measures, Table 6 presents the proportion (across assets) of estimators that are significantly worse than RV5min by class of estimator and sampling frequency.[22] Darker shaded regions represent "better" estimators, in the sense that they are rejected less often. Across the five asset classes and the entire set of assets, we observe a darker region running from the top right to the bottom left. This indicates that the simpler estimators in the top two rows (RV and RVac1) do better, on average, when implemented on lower frequency data, such as 1-minute and 5-minute data, while the more sophisticated estimators (RK, MSRV, TSRV and RRV) do relatively better when implemented on higher frequency data, such as 1-second and 5-second data.[23]

[INSERT TABLE 6 ABOUT HERE]

5.4 Estimating the *set* of best realized measures

The tests in the previous section compare a set of competing realized measures with a given benchmark measure. The RV5min measure is a reasonable, widely-used, benchmark estimator, but one might also be interested in determining whether maintaining that estimator as the "null" gives it undue preferential treatment. To address this question, we undertake an analysis based on the "model confidence set" (MCS) of Hansen et al. (2011). Given a set of competing realized measures, this approach identifies a subset that contains the unknown best estimator with some specified level of confidence, with the other measures in the MCS being not significantly different from the true best realized measure. As above, we use the QLIKE distance and a one-day lead of RVdaily

[22] In this table we aggregate across calendar-time and tick-time, trade prices and quote prices, and subsampled and not, to focus solely on the class of realized measure and sampling frequency dimensions.

[23] When Table 6 is replicated for the Romano-Wolf results obtained using the other four proxies given in Table 5, we find the same patterns. These tables are available in the online appendix. Comparing across proxies, we find that using a proxy of a certain class does not bias the Romano-Wolf results in favor of estimators of the same class.

as the proxy for QV, and Politis and Romano's (1994) stationary bootstrap with 1000 bootstrap replications and average block-size equal to 10.[24]

The number of realized measures in the model confidence sets varies across individual assets, from 3 to 143 (corresponding to a range of 1% to 37% of all measures), with the average size being 40 estimators, representing 10% of our set of 398 realized measures. By asset group, index futures and interest rate futures have the smallest model confidence sets, containing around 5% of all realized measures, and individual equities have the largest sets, containing around 17% of all measures. Table A7 in the appendix contains further information on the MCS size for each asset.

In Table 7, we summarize these results by reporting the proportion of estimators from a given class and given frequency that are included in model confidence sets, aggregating results across assets. Darker shaded elements represent the "better" realized measures. Table 7 reveals a number of interesting features. Focusing on the results for all 31 assets, presented in the upper-left panel, we see that the "best" realized measure, in terms of number of appearances in a MCS, is not 5-minute RV but 1-minute RV. Realized kernels sampled at the one-second frequency also do very well, as do TSRV and MSRV sampled at the one-second frequency.

Looking across asset classes, we see a similar pattern to that in Table 6: a dark region of good estimators includes RV and RVac1 based on lower frequency data (5 seconds to 5 minutes) and more sophisticated estimators (RK, MSRV, TSRV, MLRV and RRV) based on higher frequency data (1 second and 5 seconds). We also observe that for more liquid asset classes, such as currency futures, interest rate futures, and index futures, realized measures appear in a MCS more often if based on higher frequency data. In contrast, for individual equities and for computed equity indices, the preferred sampling frequencies are generally lower.

We can also use the estimated model confidence sets to shed light on the particularly poorly performing realized measures. Across all 31 assets, we see that realized measures based on 15-minute data almost never appear in a MCS (the only exceptions are RV and RVac1 measures for

[24]Similar to above, we also consider 15-minute RV, 5-minute RV, 1-minute MSRV, and 1-minute RKth2 as proxies for QV. Again, we find that using of one of these more accurate proxies leads to greater power in the test, i.e. smaller model confidence sets. However, the results show similar patterns to those using RVdaily as the proxy, and importantly, we find that using a proxy of a certain class (RV, TSRV, RK) does not bias the results of the test in favor of estimators of the same class. Detailed results can be found in the online appendix.

individual equities). Similarly, we observe that the more sophisticated realized measures, TSRV, MSRV, MLRV, RK and RRV are almost never in a MCS when estimated using 5-minute data: 5- and 15-minute sampling frequencies appear to be too low for these estimators. (This is consistent with the implementations of these estimators in the papers that introduced them to the literature, and so is not surprising.)

Overall, the results from the previous section revealed that it was very rare to find a realized measure that significantly outperformed 5-minute RV. The analysis in this section, which avoids the need to specify a "benchmark" realized measure, reveals evidence that some measures are indeed more accurate than 5-minute RV. We find that 1-minute RV and RVac1, 1-second and 5-second realized kernels and multi-scale RV, and 5-second and 1-minute realized range estimators appear more often in the MCS than 5-minute RV.

[INSERT TABLE 7 ABOUT HERE]

5.5 Variations in accuracy

The Romano-Wolf tests and model confidence sets investigate average accuracy over the sample period, from 2000 to 2010. These 11 years contain several subperiods during which asset volatility and market behavior were very different, and by conducting tests over the entire period we may miss some significant differences in *conditional* accuracy that are averaged out over the full sample.

To investigate this further, we implement tests of relative conditional accuracy using the approach of Giacomini and White (2006). This approach can be used to study whether the relative performance of two realized measures varies with some conditioning variable, Z. We consider two conditioning variables: volatility, measured using the log-average RVdaily for the asset over the previous 10 trading days, and liquidity, measured using the average log-spread for the asset over the past 10 trading days. We estimate regressions that compare RV5min with a few of the better performing realized measures identified in the previous section, namely, 5-second MSRV, 1-minute RVac1, and 5-second RKth2.[25] We also include 1-minute RV and RVdaily to study the accuracy

[25] The fact that we examine realized measures identified as "good" in previous analysis of course biases the interpretation of any subsequent tests of *unconditional* accuracy. In this section we focus on whether the relative performance

gains from using higher-frequency price data. All of these estimators are computed on transaction prices with calendar-time sampling. We estimate this model using an unbalanced panel framework, allowing for different unconditional relative accuracy across assets, but imposing a common coefficient on the conditioning variable. For a given pair of realized measures $(M^i_{0,t}, M^i_{j,t})$, we estimate:

$$L(\tilde{\theta}^i_t, M^i_{0,t}) - L(\tilde{\theta}^i_t, M^i_{j,t}) = \alpha_{i,j} + \beta_j Z^i_{t-1}, \quad \text{for } t = 1, 2, ...T;\ i = 1, 2, ..., 31 \tag{7}$$

where $\tilde{\theta}^i_t$ is the volatility proxy, a lead of RVdaily. A positive value of β_j indicates that higher values of Z lead to an improvement in the performance of the alternative realized measure, $M^i_{j,t}$, relative to $M^i_{0,t}$ =RV5min. We estimate this panel model for all 31 assets jointly, and also for subpanels comprising of assets from a single class.

The t-statistics for the coefficient on Z from the panel regressions are presented in Table 8. For daily squared returns we see that all coefficients on volatility are negative and strongly significant for all but the class of currency futures. This reveals that daily squared returns, which are significantly worse than RV5min unconditionally, perform even worse when volatility is high. We find a similar result for MSRV, RK and 1-minute RV, with their relative performance declining in highly volatile markets, however these results are both driven purely by the set of computed indices, which is the set where the MSRV, RK and 1-minute RV measures did not perform well unconditionally.

Using recent liquidity, measured via the bid-ask spread, we find that the relative performance of MSRV and the 1-minute RV estimators compared to RV5min declines as spreads increase (i.e., as liquidity decreases). For both of these realized measures, this is true when using all assets, and is driven by significant results for the class of individual equities and index futures. The performances of RK and RVac1, on the other hand, do not appear to be significantly affected by changes in market liquidity.

[INSERT TABLE 8 ABOUT HERE]

of these measures varies significantly with some conditioning variable Z, and the problem of pre-test bias does not arise here.

5.6 Out-of-sample forecasting with realized measures

The results above have all focussed on the relative accuracy of realized measures for *estimating* quadratic variation. One of the main uses of estimators of volatility is in the production of volatility forecasts, and in this section we compare the relative accuracy of forecasts based on our set of competing realized measures. We do so based on the simple heterogeneous autoregressive (HAR) forecasting model of Corsi (2009). This model is popular in practice as it captures long memory-type properties of quadratic variation, while being simpler to estimate than fractionally integrated processes, and performs well in volatility forecasting, see Andersen et al. (2007) for example.[26] For each realized measure, we estimate the HAR model using the most recent 500 days of data:

$$\tilde{\theta}_{t+h} = \beta_{0,j,h} + \beta_{1,j,h} M_{jt} + \beta_{2,j,h} \frac{1}{5} \sum_{k=0}^{4} M_{j,t-k} + \beta_{3,j,h} \frac{1}{22} \sum_{k=0}^{21} M_{j,t-k} + \varepsilon_{jt}, \tag{8}$$

where M_{jt} is a realized measure from the competing set, and $\tilde{\theta}_{t+h}$ is the volatility proxy (the squared open-to-close return for day $t+h$). We estimate this regression separately for each forecast horizon, h, ranging from 1 to 50 trading days, and from those estimates we obtain a h-day ahead volatility forecast, which we then compare with our volatility proxy. We re-estimate the model each day using a rolling window of 500 days.

In addition to the 398 realized measures we have analyzed so far, for forecasting analysis we also consider some "jump-robust" estimators of volatility. These measures, described in Section 2.3, are designed to estimate only the integrated variance component of quadratic variation, see equation 2. The inclusion of these estimators is motivated by studies such as Andersen et al. (2007) and Patton and Sheppard (2011) which report that the predictability of the integrated variance component of quadratic variation is stronger than the jump component, and thus there may be gains to separately forecasting the two components. Using a HAR model on these jump-robust realized measures effectively treats the jump component as unpredictable, while using a HAR model on estimators of QV (our original set of 398 measures) treats the two components as having equal

[26] Alternatives to this specification include a simple AR(1), as used by Aït-Sahalia and Mancini (2008), an AR(p), as in Andersen et al. (2011), or a MIDAS regression, as in Ghysels and Sinko (2011). These models are all similar in structure and the results we obtain below are unlikely to differ greatly across these choices of forecasting model.

predictability. Extending our set to include 206 jump-robust measures increases its total number to 604 realized measures.

For each forecast horizon between one day and 50 days we estimate the model confidence set of Hansen et al. (2011). It is not feasible to report the results of each of these estimates for each horizon, and so we summarize them in two ways. Firstly, in Figure 1 below we present the size of the MCS, measured as the proportion of realized measures that are included in the MCS, across forecast horizons. From this figure we observe that the MCSs are relatively small for short horizons, consistent with our results in Section 5.4 and with the well-known strong persistence in volatility. As the forecast horizon grows, the size of the MCSs increase, reflecting the fact that for longer horizons more precise measurement of current volatility provides less of a gain than for short horizons. It is noteworthy that even at horizons of 50 days, we are able to exclude around 35% of realized measures from the MCS, averaging across all 31 assets. This proportion varies across asset classes, with the proportion of estimators included at $h = 50$ being around 22% for the liquid class of interest rate futures, and being almost 100% (i.e., no realized measures are excluded) for the illiquid class of computed equity indices.

[INSERT FIGURE 1 ABOUT HERE]

In Table 9 we study these results in greater detail. This table has the same format as Table 7, and reports the proportion of realized measures from a given class and given frequency that belong to a model confidence set, aggregating results across assets and forecast horizons between 1 and 5 days. As in Table 7, darker shaded elements represent the better realized measures. What is most striking about this table is the relative success of the jump-robust realized measures for volatility forecasting. For each of the 5 asset classes, the best measure is one of truncated-RV (TRV) at the 5-minute or 15-minute frequency, or quantile-RV at the 5-minute frequency. This pattern is consistent across all asset classes: the best realized measures for volatility forecasting appear to be jump-robust measures, estimated using relatively low (5- or 15-minute) frequency data.

[INSERT TABLE 9 ABOUT HERE]

In Figure 2 below we present the proportion (across assets) of model confidence sets that contain RV5min and TRV5min (both computed on trades prices with calendar-time sampling), for each forecast horizon. We see that, across all assets, RV5min appears in around 40% of MCSs for shorter horizons, rising to around 70% for longer horizons.[27] RV5min does best for currency futures, equity index futures and computed indices, and relatively poorly for interest rate futures. Figure 2 also presents the corresponding proportion for truncated RV5min, and we see that this measure does almost uniformly better than RV5min, with the exceptions being for the individual equities and index futures on longer forecast horizons. TRV5min does particularly well for currency futures and interest rate futures.

[INSERT FIGURE 2 ABOUT HERE]

Our study of a broad collection of assets and a large set of realized measures necessitates simplifying the analysis in several ways, and a few caveats to the above conclusions apply. Firstly, these results are based on each realized measure being used in conjunction with the HAR model of Corsi (2009). This model has proven successful in a variety of volatility applications, but it is by no means the only relevant volatility forecasting model in the literature, and it is possible that the results and rankings change with the use of a different model. Secondly, by treating the prediction of future QV as a univariate problem, we have implicitly made a strong assumption about the predictability of volatility attributable to jumps, either that it is identical to that of integrated variance, or that it is not predictable at all. A more sophisticated approach might treat these two components separately. Thirdly, we have only considered forecasting models based on a single realized measure, and it may be possible that a given realized measure is not very useful on its own, but informative when combined with another realized measure.

[27] Note that this analysis only counts RV5min computed in calendar time, using transaction prices, and not sub-sampled. Thus this represents a lower bound on the proportion of MCSs that include *any* RV5min.

6 Summary and conclusion

Motivated by the large body of research on estimators of asset price volatility using high frequency data (so-called "realized measures"), this paper considers the problem of comparing the empirical accuracy of a large collection these measures across a range of assets. In total, we consider almost 400 different estimators, applied to 11 years of data on 31 different financial assets across five asset classes, including equities, indices, exchange rates and interest rates. We apply data-based ranking methods to the realized measures and to forecasts based on these measures, for forecast horizons ranging from 1 to 50 trading days.

Our main findings can be summarized as follows. Firstly, if 5-minute RV is taken as the benchmark realized measure, then using the testing approach of Romano and Wolf (2005) we find very little evidence that it is significantly outperformed by *any* of the competing measures, in terms of estimation accuracy, across any of the 31 assets under analysis. If, on the other hand, the researcher wishes to remain agnostic about the "benchmark" realized measure, then using the model confidence set of Hansen et al. (2011), we find that 5-minute RV is indeed outperformed by a small number of estimators, most notably 1-minute RV and RVac1, and 1- and 5-second realized kernels and MSRV. Finally, when using forecast performance as the method of ranking realized measures, we find that 5-minute or 15-minute truncated RV provides the best performance on average, which is consistent with the work of Andersen et al. (2007), who find that jumps are not very persistent. The rankings of realized measures vary across asset classes, with 5-minute RV performing better on the relatively less liquid classes (individual equities and computed equity indices), and the gains from more sophisticated estimators like MSRV and realized kernels being more apparent for more liquid asset classes (such as currency futures and equity index futures). We also find that for realized measures based on frequencies of around five minutes, sampling in tick time and subsampling the realized measure both generally lead to increased accuracy.

References

Aït-Sahalia, Y. and Mancini, L. (2008). Out of sample forecasts of quadratic variation. *Journal of Econometrics*, 147:17–33.

Aït-Sahalia, Y., Mykland, P., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies*, 18:351–416.

Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *Journal of Finance*, 57:1047–1092.

Andersen, T., Bollerslev, T., Diebold, F., and Ebens, H. (2001a). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76.

Andersen, T., Bollerslev, T., Diebold, F., and Labys, P. (2001b). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, pages 42–55.

Andersen, T., Bollerslev, T., and Meddahi, N. (2005). Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. *Econometrica*, 73(1):279–296.

Andersen, T., Bollerslev, T., and Meddahi, N. (2011). Realized volatility forecasting and market microstructure noise. *Journal of Econometrics*, 160(1):220–234.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39:885–905.

Andersen, T. G., Bollerslev, T., Christoffersen, P., and Diebold., F. X. (2006). Volatility and correlation forecasting. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting, Volume 1.* Elsevier, Oxford.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling and forecasting of return volatility. *Review of Economics and Statistics*, 89:701–720.

Andersen, T. G., Dobrev, D., and Schaumburg, E. (2008). Jump robust volatility estimation using nearest neighbour truncation. *Journal of Econometrics*. Forthcoming.

Bandi, F. and Russell, J. (2008). Microstructure noise, realized variance, and optimal sampling. *Review of Economic Studies*, 75(2):339–369.

Barndorff-Nielsen, O. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280.

Barndorff-Nielsen, O., Hansen, P., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O., Hansen, P., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *The Econometrics Journal*, 12(3):C1–C32.

Barndorff-Nielsen, O., Hansen, P., Lunde, A., and Shephard, N. (2011). Subsampling realised kernels. *Journal of Econometrics*, 160(1):204–219.

Barndorff-Nielsen, O. and Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4(1):1–30.

Barndorff-Nielsen, O. E. and Shephard, N. (2007). Variation, jumps, market frictions and high frequency data in financial econometrics. In Blundell, R., Torsten, P., and Newey, W. K., editors, *Advances in economics and econometrics. Theory and applications*, Econometric Society monographs, pages 328–372. Cambridge University Press, Cambridge.

Bollerslev, T., Engle, R., and Nelson, D. (1994). ARCH models. In *Handbook of Econometrics*, pages 2959–3038. Elsevier.

Christensen, K., Oomen, R., and Podolskij, M. (2010). Realised quantile-based estimation of the integrated variance. *Journal of Econometrics*, 159(1):74–98.

Christensen, K. and Podolskij, M. (2007). Realized range-based estimation of integrated variance. *Journal of Econometrics*, 141(2):323–349.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Diebold, F. and Mariano, R. (2002). Comparing predictive accuracy. *Journal of business and economic statistics*, 20(1):134–144.

French, K., Schwert, G., and Stambaugh, R. (1987). Expected stock returns and volatility. *Journal of financial Economics*, 19(1):3–29.

Gatheral, J. and Oomen, R. (2010). Zero-intelligence realized variance estimation. *Finance and Stochastics*, 14:249–283.

Ghysels, E. and Sinko, A. (2011). Volatility forecasting and microstructure noise. *Journal of Econometrics*, 160(1):257271.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Hansen, P. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics*, 23(4):365–380.

Hansen, P. and Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7):873–889.

Hansen, P. and Lunde, A. (2006a). Consistent ranking of volatility models. *Journal of Econometrics*, 131(1-2):97–121.

Hansen, P. and Lunde, A. (2006b). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics*, 24(2):127–161.

Hansen, P. and Lunde, A. (2010). Estimating the persistence and the autocorrelation function of a time series that is measured with error. *Manuscript, Stanford University and University of Aarhus.*

Hansen, P., Lunde, A., and Nason, J. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hansen, P. R. and Lunde, A. (2006c). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics*, 24:127–161.

Mancini, C. (2001). Disentangling the jumps of the diffusion in a geometric jumping brownian motion. *Giornale dellIstituto Italiano degli Attuari*, 64:19–47.

Mancini, C. (2009). Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 36(2):270–296.

Martens, M. and Van Dijk, D. (2007). Measuring volatility with the realized range. *Journal of Econometrics*, 138(1):181–207.

Meddahi, N., Mykland, P., and Shephard, N. (2011). Special issue on realised volatility. *Journal of Econometrics*, 160.

Oomen, R. C. A. (2006). Properties of realized variance under alternative sampling schemes. *Journal of Business and Economic Statistics*, 24:219–237.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, 53:61–65.

Patton, A. (2011a). Data-based ranking of realised volatility estimators. *Journal of Econometrics*, 161(2):284–303.

Patton, A. (2011b). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Patton, A. and Sheppard, K. (2009a). Optimal combinations of realised volatility estimators. *International Journal of Forecasting*, 25(2):218–238.

Patton, A. J. and Sheppard, K. (2009b). Evaluating volatility and correlation forecasts. In Andersen, T. G., Davis, R. A., Kreiss, J.-P., and Mikosch, T., editors, *Handbook of Financial Time Series*. Springer, Verlag.

Patton, A. J. and Sheppard, K. (2011). Good volatility, bad volatility: Signed jumps and the persistence of volatility. Working paper, Duke University.

Politis, D. and Romano, J. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, pages 1303–1313.

Romano, J. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 12(6):1019–1043.

Zhang, L., Mykland, P., and Ait-Sahalia, Y. (2005). A tale of two time scales. *Journal of the American Statistical Association*, 100(472):1394–1411.

Zhou, B. (1996). High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics*, pages 45–52.

Table 1
Decription of Price Data

Assets		Dates		T	Avg. Ann. Vol	Avg. Trade Dur.	Avg. Quote Dur.
U.S. Equities							
KO	Coca-Cola	1/3/2000	12/31/2010	2766	18.8	7.6	2.6
SYY	Sysco	1/3/2000	12/31/2010	2766	22.1	12.5	3.4
IFF	Intl. Flavors & Fragrances	1/3/2000	12/31/2010	2767	23.9	26.6	5.4
MSFT	Microsoft	1/3/2000	12/31/2010	2763	24.5	2.7	1.5
LSI	LSI corp.	1/3/2000	12/31/2010	2767	48.5	15.6	3.8
U.K. Equities							
DGE	Diageo	1/4/2000	12/31/2010	2769	23.9	15.8	3.6
SAB	SABMiller	1/4/2000	12/31/2010	2733	27.9	23.6	3.8
VOD	Vodaphone	1/4/2000	12/31/2010	2770	29.5	7.0	2.3
RSA	RSA Ins.	1/4/2000	12/31/2010	2768	39.1	28.1	6.4
SDR	Schroders	1/4/2000	12/31/2010	2757	45.8	52.4	8.7
Index futures							
JNI	Nikkei 225	1/4/2000	10/29/2010	2644	15.2	3.5	0.9
ES	e-mini S&P 500	1/3/2000	12/31/2010	2750	14.6	0.5	0.2
FFI	FTSE 100	1/4/2000	10/29/2010	2707	15.6	1.9	0.5
STXE	EuroStoxx50	1/3/2000	12/30/2010	2782	17.9	2.0	0.7
FDX	DAX 40	1/3/2000	10/29/2010	2738	17.9	1.5	0.8
Interest Rate Futures							
TU	2 yr Treasury note	1/2/2003	12/31/2010	1994	1.4	7.6	0.5
FV	5 yr Treasury note	1/2/2001	12/31/2010	2486	3.5	3.0	0.3
TY	10 yr Treasury note	1/2/2001	12/31/2010	2484	5.2	1.9	0.3
US	30 yr Treasury bond	1/2/2001	10/29/2010	2449	8.1	2.4	0.4
FGBS	German short term govt bond	1/3/2000	10/29/2010	2735	1.3	9.0	1.9
FGBL	German long term govt bond	1/3/2000	10/29/2010	2741	4.6	2.7	1.0
Currency futures							
BP	British Pound	1/2/2004	12/31/2010	1762	6.7	2.9	0.4
URO	Euro	1/2/2004	12/31/2010	1762	6.9	1.4	0.3
JY	Japanese Yen	1/2/2004	12/31/2010	1763	7.3	3.1	0.4
CD	Canadian Dollar	1/2/2004	12/31/2010	1763	8.4	4.1	0.6
AD	Australian Dollar	1/2/2004	12/30/2010	1759	9.3	4.9	0.5
Market Indices							
N225	Nikkei 225	1/5/2000	12/30/2010	2665	14.7	48.1	-
SPX	S&P500	1/3/2000	12/31/2010	2719	16.1	15.9	-
FTSE	FTSE 100	1/4/2000	12/31/2010	2762	15.9	4.9	-
STOXX50E	EuroStoxx50	1/3/2000	12/30/2010	2782	18.6	15.2	-
DAX	DAX 40	1/4/2006	12/30/2010	2781	19.4	2.9	-

Notes: This table presents the 31 assets included in the analysis, the sample period for each asset, and some summary statistics: the number of observations, the average volatility (annualized, estimated using squared open-to-close returns), the average trade duration (in seconds) and the average quote duration.

Table 2
Summary Statistics of some sample realized measures for two representative assets

	mean	std dev	skew	kurt	min	max	rho(1)	rho(2)	rho*(1)	rho*(2)
					Microsoft (MSFT) returns					
RVdaily	3.20	2.69	6.53	72.09	0.00	112.86	0.26	0.29	0.96	0.99
RV_5m_ct	3.37	2.12	4.56	36.86	0.18	63.14	0.72	0.68	0.96	0.95
RV_5m_ct_ss	2.73	1.97	4.75	40.05	0.07	54.96	0.65	0.62	0.97	0.95
RV_1t_bt	11.24	4.51	3.75	20.96	0.27	207.58	0.94	0.92	0.99	0.98
RVac1_1m_ct	3.40	2.13	5.22	53.70	0.15	81.89	0.72	0.70	0.94	0.94
RKth2_1m_bt	3.19	2.11	4.76	40.18	0.13	66.49	0.70	0.65	0.96	0.95
MSRV_1m_ct	3.23	2.12	4.81	41.16	0.13	68.19	0.69	0.65	0.96	0.95
MLRV_5s_ct	3.21	3.62	5.02	50.41	0.26	63.32	0.80	0.77	0.95	0.93
RRVm5_1m_bt	3.34	2.06	5.37	61.72	0.21	81.49	0.74	0.72	0.94	0.93
					USD/AUD exchange rate (AD) returns					
RVdaily	0.46	1.17	9.88	149.55	0.00	28.95	0.39	0.40	0.98	0.93
RV_5m_ct	0.52	1.02	7.90	91.46	0.04	17.21	0.71	0.78	0.94	0.93
RV_5m_ct_ss	0.49	1.05	9.29	125.09	0.02	19.56	0.67	0.75	0.86	0.85
RV_1t_bt	0.70	1.02	7.61	92.73	0.07	18.37	0.70	0.70	0.95	0.91
RVac1_1m_ct	0.52	1.01	7.95	96.27	0.04	18.14	0.73	0.78	0.94	0.93
RKth2_1m_bt	0.50	1.01	8.04	94.36	0.04	16.31	0.71	0.78	0.91	0.90
MSRV_1m_ct	0.51	1.02	8.06	95.30	0.04	17.04	0.72	0.79	0.92	0.91
MLRV_5s_ct	0.57	0.99	6.91	71.92	0.06	16.06	0.79	0.78	0.96	0.92
RRVm5_1m_bt	0.54	1.00	7.29	78.92	0.05	16.25	0.78	0.79	0.95	0.91

Note: This table displays the summary statistics for several estimators for Microsoft and Australian-US Dollar futures. Referring to the four right-most columns, 'rho' denotes the sample autocorrelation, and 'rho*' denotes the estimated autocorrelation of QV based on a realized measure, using the instrumental variables method of Hansen and Lunde (2010).

Table 3
Summary of the top 10 estimators across each asset class

		Indiv. Equities	**avg rank**		**Bond Futures**	**avg rank**		**Currency Futures**	**avg rank**		**Index Futures**	**avg rank**	**Comp. Index**	**avg rank**
	mq	RKth2_5s_b	31	mq	RRVm5_5s_b	28	tr	TSRV_1s_c_ss	20	tr	RV_1m_b_ss	33	RVac1_1m_b	7
	mq	RKbart_5s_b	33	mq	RRVm5_5s_b_ss	29	tr	TSRV_1s_c	21	tr	RVac1_1m_b_ss	33	RVac1_1m_c	8
	mq	RKnfp_1s_b	35	mq	RRVm10_1s_c_ss	34	mq	MSRV_1s_b_ss	23	tr	RV_1m_b	36	RKth2_1t_b	11
	mq	RKbart_1s_b	38	mq	RRVm10_1s_c	34	mq	MLRV_1s_c	24	tr	MSRV_5s_c_ss	36	RKcub_1t_b	12
	tr	RKnfp_1s_b	38	mq	RRVm10_1s_b_ss	35	mq	MLRV_1s_c_ss	25	tr	RKbart_1s_c	37	RKbart_1m_b	13
	tr	RRVm10_1m_b_ss	39	mq	RRVm5_5s_c_ss	36	mq	MSRV_1s_b	25	tr	MSRV_5s_c	38	RKbart_1t_b	15
	mq	RKcub_1s_b	43	mq	RRVm10_1s_b	36	mq	RV_5s_c	26	tr	RKth2_1s_c	40	RKth2_1m_b	15
	tr	RRVm10_1m_b	43	mq	RRVm5_5s_c	36	mq	RV_5s_c_ss	27	tr	RVac1_1m_c_ss	40	RKbart_1m_c	15
	mq	RV_1m_c_ss	44	tr	RKth2_1s_c	37	mq	MSRV_1s_c	28	tr	RKbart_1s_b	41	RKnfp_1t_b	15
	mq	RKbart_1s_c	48	mq	RVac1_5s_c	38	mq	MSRV_1s_c_ss	28	tr	RRVm10_5s_b_ss	42	RKnfp_5s_b	17

Asset Class	**Freq.**	**No.**	**Freq.**	**No.**	**Freq.**	**No.**	**Freq.**	**No.**	**Freq.**	**No.**
RV	1m	1	-	-	5s	2	1m	2	-	-
RVac1	-	-	5s	1	-	-	1m	2	1m	2
RK										
bart	1s,5s	3	-	-	-	-	1s	2	1t, 1m	3
cub	1s	1	-	-	-	-			1t	1
th2	5s	1	1s	1	-	-	1s	1	1t, 1m	2
nfp	1s	2	-	-	-	-	-	-	1t, 5s	2
kSRV										
tsrv	-	-	-	-	1s	2	-	-	-	-
msrv	-	-	-	-	1s	4	5s	2	-	-
MLRV	-	-	-	-	1s	2	-	-	-	-
RRV										
rrv5	-	-	5s	4	-	-	-	-	-	-
rrv10	1m	2	1s	4	-	-	5s	1	-	-

Notes: For each asset class, we take the average of the rankings from all assets of that class. The top panel of this table lists the estimators with the top "average-ranks" for each asset class. The bottom panel summarizes the top panel by categorizing them by estimator characteristics.

Table 4.
Pairwise comparisons of estimators

Calendar-time sampling vs Tick-time sampling

	1s	5s	1m	5m	15m
RV	-84	-74	0	23	35
RVac1	-84	-68	-3	42	29
RK	-13	6	48	48	39
MSRV	-45	-19	29	42	32
TSRV	35	-52	-3	32	42
MLRV	-81	-45	6	42	19
RRV	-61	3	55	77	81
BR	0				

Transaction prices vs Mid-quote prices

	1t	1s	5s	1m	5m	15m
RV	73	73	-19	-62	-81	-81
RVac1	-38	-4	-38	-54	-42	-42
RK	-4	27	-15	-65	-96	-88
MSRV	-42	-23	-4	77	50	-23
TSRV	-65	-92	-42	-12	27	4
MLRV	-31	77	23	-50	-46	4
RRV	-23	69	19	23	0	23
BR	8					

Not subsampled vs Subsampled estimators

	1s	5s	1m	5m	15m
RV	3	6	6	29	52
RVac1	-58	-39	29	84	94
MSRV	-3	0	10	19	0
TSRV	0	0	16	55	35
MLRV	0	3	6	65	77
RRV	0	-26	-19	-42	-58
BR	6				

Note: This table presents results on comparisons of realized measures that differ only in the sampling scheme used (top panel), price series used (middle panel), or use of subsampling (lower panel). For each pair of measures, a robust t-statistic on the averae difference in loss is computed. Each element of the table presents the proportion of significantly positive t-statistics minus the proportion of significantly negative t-statistics. A negative value indicates that the first approach (e.g., calendar-time sampling in the top panel) outperforms the second approach, a positive value indicates the opposite. Elements with values less than -33 are dark-shaded; those with values greater than 33 are light-shaded.

Table 5.
Number of estimators that are significantly different from RV5min in Romano-Wolf Tests

	Worse					Better					Total Estimators
Proxy:	*RV Daily*	*RV 15min*	*RV 1min*	*MSRV 1min*	*RKth2 1min*	*RV Daily*	*RV 15min*	*RV 1min*	*MSRV 1min*	*RKth2 1min*	
KO	161	231	219	240	237	0	0	0	0	0	396
LSI	160	265	257	272	278	0	0	0	0	0	395
MSFT	243	285	272	288	290	0	0	0	0	0	396
IFF	127	238	254	259	252	0	0	0	0	0	391
SYY	129	210	206	190	190	0	0	0	0	0	392
DGE	157	318	335	231	247	0	0	0	0	0	398
VOD	179	279	351	211	212	0	0	0	0	0	397
SAB	126	322	278	312	316	0	0	0	0	0	398
SDR	116	301	295	274	277	0	0	0	0	0	394
RSA	141	291	362	165	202	0	0	0	0	0	397
TU	204	180	194	166	187	0	0	0	0	0	397
FV	192	237	220	221	236	0	0	0	0	0	398
TY	188	229	213	211	225	0	9	24	28	23	398
US	202	247	241	243	254	0	0	0	0	0	397
FGBL	183	269	266	267	268	0	0	0	0	0	398
FGBS	310	367	131	363	343	0	0	0	0	0	398
CD	120	177	178	177	178	0	0	0	0	0	398
AD	102	171	173	180	181	0	0	0	0	0	398
BP	134	166	170	165	166	0	0	0	0	0	398
URO	149	167	172	172	172	0	0	0	0	0	398
JY	139	172	178	175	172	0	0	0	0	0	398
STXE	177	60	183	280	284	0	0	0	0	0	398
JNI	250	324	331	317	318	0	0	0	0	0	394
FDX	142	145	145	182	181	0	0	0	0	0	398
FFI	150	183	182	184	185	0	0	0	0	0	398
ES	159	204	204	204	206	0	0	0	0	0	398
SPX	156	169	169	155	163	0	0	0	7	1	199
STOXX50E	123	170	168	143	166	0	0	0	0	0	199
DAX	122	148	155	147	152	0	0	0	0	0	199
FTSE	153	175	172	129	169	0	0	0	0	0	199
N225	143	159	161	161	160	0	0	0	0	0	197

Note: Results from when a potential proxy has significantly different mean from RVdaily are displayed in lighter color.

Table 6.
Percentage of estimators that are significantly worse than RV5min

	All 31 Assets						*Currency Futures*						*Interest Rate Futures*					
	1t	**1s**	**5s**	**1m**	**5m**	**15m**	**1t**	**1s**	**5s**	**1m**	**5m**	**15m**	**1t**	**1s**	**5s**	**1m**	**5m**	**15m**
RV	70	55	39	18	17	71	70	36	10	0	0	65	75	59	33	8	52	100
RVac1	30	41	27	19	49	73	0	6	0	0	40	58	36	41	19	46	81	96
RK	11	15	18	50	87	91	0	0	1	41	93	88	40	18	55	98	99	97
MSRV	21	24	13	43	93	87	0	0	0	25	98	78	25	15	8	96	98	89
TSRV	75	39	71	97	98	96	60	24	70	100	98	100	58	15	58	100	100	94
MLRV	28	38	22	22	84	78	0	6	0	0	80	80	33	34	17	83	100	85
RRV	25	35	27	22	66	95	0	0	0	1	58	100	13	19	13	50	100	98
BR	18						0						31					

	Individual Equities						*Index Futures*						*Computed Indices*					
	1t	**1s**	**5s**	**1m**	**5m**	**15m**	**1t**	**1s**	**5s**	**1m**	**5m**	**15m**	**1t**	**1s**	**5s**	**1m**	**5m**	**15m**
RV	65	62	54	25	0	40	60	44	30	0	23	95	100	100	100	80	20	85
RVac1	40	58	39	14	29	60	10	33	15	10	55	83	80	100	100	30	60	85
RK	0	14	3	28	71	86	10	3	10	64	100	95	5	65	45	18	80	98
MSRV	20	44	17	13	84	84	20	8	5	58	98	93	60	100	71	50	95	94
TSRV	80	49	61	91	96	92	90	49	95	100	100	98	100	100	100	100	100	100
MLRV	30	58	32	0	71	61	20	28	5	15	93	95	80	100	100	20	85	85
RRV	25	50	37	5	39	86	25	28	15	16	79	100	100	100	100	75	80	100
BR	11						5						75					

Note: This table aggregates, for groups of assets (either all 31 assets or assets belonging to one class), the Romano-Wolf test results identifying estimators that are significantly worse than the benchmark 5-minute RV (calendar-time, trades prices) estimator. Each table cell reports the proportion of estimators of a certain estimator class and sampling frequency (across assets, and allowing for different sampling schemes and sampled price series) that are found to be significantly worse than the benchmark estimator in a Romano-Wolf test.

Table 7
Percentage of estimators that are in a 90% MCS

	All 31 Assets						*Currency Futures*						*Interest Rate Futures*					
	1t	**1s**	**5s**	**1m**	**5m**	**15m**	**1t**	**1s**	**5s**	**1m**	**5m**	**15m**	**1t**	**1s**	**5s**	**1m**	**5m**	**15m**
RV	4	3	15	30	18	1	0	3	25	23	15	0	0	0	0	21	8	0
RVac1	7	6	18	27	8	0	0	15	33	15	8	0	0	0	8	8	2	0
RK	18	29	26	6	0	0	10	21	19	1	0	0	0	10	2	0	0	0
MSRV	9	24	21	3	0	0	0	39	15	5	0	0	0	24	21	0	0	0
TSRV	0	11	4	0	0	0	0	12	0	0	0	0	0	22	4	0	0	0
MLRV	9	15	22	15	0	0	0	36	35	10	0	0	0	0	23	0	0	0
RRV	15	11	17	20	2	0	5	20	18	15	0	0	4	10	22	6	0	0
BR	9						14						0					

	Individual Equities						*Index Futures*						*Computed Indices*					
	1t	**1s**	**5s**	**1m**	**5m**	**15m**	**1t**	**1s**	**5s**	**1m**	**5m**	**15m**	**1t**	**1s**	**5s**	**1m**	**5m**	**15m**
RV	10	9	13	49	33	3	0	0	30	25	0	0	0	0	0	5	25	0
RVac1	20	11	19	46	16	1	0	3	20	10	0	0	0	0	0	50	5	0
RK	34	58	54	4	0	0	0	10	3	0	0	0	50	10	28	48	0	0
MSRV	20	18	38	0	0	0	0	26	8	0	0	0	20	0	0	20	0	0
TSRV	0	11	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MLRV	25	11	22	29	0	0	0	21	18	0	0	0	0	0	0	40	0	0
RRV	30	11	18	45	5	0	15	8	19	0	0	0	0	0	0	0	5	0
BR	17						5						0					

Note: This table aggregates, for groups of assets (either all 31 assets or assets belonging to one class), the 90% Model Confidence Sets identifying the subset containing "best" estimators. Each table cell reports the percentage of all estimators of a certain estimator class and sampling frequency (across assets, and aggregating estimators using different sampling schemes and sampled price series) that are found to be in a Model Confidence Set.

Table 8
Impact of volatility or liquidity on the relative performance of realized measures

"Other" Estimator:	*RV_daily*	*RV_1m*	*RVac1_1m*	*MSRV_5s*	*RKth2_5s*
			t-stats on lagged volatility		
All assets	**-5.71**	-1.54	**3.46**	**-3.87**	-1.84
Individual Equities	**-3.08**	**2.68**	0.87	1.03	1.18
Interest Rate Futures	**-2.00**	-1.27	**4.69**	-1.23	-0.61
Currency Futures	-1.52	-0.98	-0.09	-0.93	-0.73
Index Futures	**-3.75**	-0.73	1.26	-1.89	-1.71
Computed Indices	**-4.91**	**-3.48**	-0.09	**-4.62**	**-2.39**
			t-stats on lagged liquidity		
All assets	-0.73	**-3.49**	-1.23	**-2.43**	-0.91
Individual Equities	0.34	**-6.07**	-1.53	**-4.10**	-1.53
Interest Rate Futures	**3.22**	0.57	-0.68	0.59	1.21
Currency Futures	-1.41	-0.62	0.32	-0.79	-0.45
Index Futures	**-3.47**	**-2.47**	-0.28	**-2.21**	-1.46

Note: This table presents β estimates from the panel regression $\text{Loss}(\text{RV5min}_i) - \text{Loss}(\text{Other}_i) = \alpha_i + \beta Z_i + \epsilon_i$, where Z_i is the lagged 10-day average of 'volatility' as measured by daily squared returns, or 'liquidity' measured by the mean log(ask)–log(bid) over a day. All estimators are calendar-time sampled, transaction price estimators. Quote data for computed indices is not available, and so this asset class is not reported in the lower panel. Statistically significant results (at 5% level) are bolded.

Table 9
Percentage of RM-based HAR-RV models that belong to the 90% MCS for forecast horizons 1 through 5

All Assets

	1t	1s	5s	1m	5m	15m
RV	11	14	19	29	38	41
RVac1	15	17	19	31	36	30
RK	36	32	36	40	35	17
MSRV	29	23	25	29	29	12
TSRV	11	14	18	32	36	15
MLRV	21	18	23	32	37	14
RRV	18	16	20	39	43	31
BR	38					
BPV	9	10	15	28	40	49
minRV	9	10	15	26	34	42
medRV	10	9	14	25	40	47
QRV	10	9	17	36	64	54
TrunRV	14	3	17	43	63	66

Interest Rate Futures

	1t	1s	5s	1m	5m	15m
RV	2	5	9	1	7	8
RVac1	2	8	5	2	8	8
RK	6	2	3	9	11	8
MSRV	15	12	5	7	13	11
TSRV	0	7	6	4	10	12
MLRV	5	8	8	3	10	13
RRV	1	6	2	5	15	22
BR	8					
BPV	0	0	0	7	23	37
minRV	0	0	1	10	20	23
medRV	0	0	3	8	23	31
QRV	0	1	9	38	71	63
TrunRV	10	-	10	59	78	81

Index Futures

	1t	1s	5s	1m	5m	15m
RV	14	12	15	20	23	26
RVac1	6	14	10	21	25	13
RK	31	21	26	23	20	2
MSRV	42	16	9	11	11	0
TSRV	10	2	6	21	25	1
MLRV	24	14	12	23	21	0
RRV	22	14	12	24	18	4
BR	52					
BPV	2	3	25	21	27	30
minRV	6	5	23	18	18	28
medRV	4	5	17	16	27	31
QRV	4	13	14	33	49	27
TrunRV	0	0	11	41	60	46

Individual Equities

	1t	1s	5s	1m	5m	15m
RV	10	10	13	31	48	47
RVac1	19	12	18	37	38	38
RK	43	38	45	51	40	19
MSRV	27	22	29	30	32	7
TSRV	13	14	22	38	33	15
MLRV	21	12	20	38	39	12
RRV	18	11	13	44	56	35
BR	39					
BPV	8	6	7	21	33	46
minRV	9	6	7	21	28	38
medRV	8	8	7	21	34	44
QRV	8	0	8	14	51	49
TrunRV	11	4	7	15	37	52

Currency Futures

	1t	1s	5s	1m	5m	15m
RV	10	26	38	65	76	80
RVac1	22	35	40	63	79	54
RK	67	66	72	80	76	45
MSRV	28	43	55	72	59	37
TSRV	14	30	34	61	79	38
MLRV	26	42	44	65	79	42
RRV	30	35	56	81	80	67
BR	58					
BPV	16	29	34	73	85	89
minRV	18	25	33	68	76	85
medRV	22	20	33	69	79	88
QRV	20	20	45	84	91	91
TrunRV	20	5	42	90	98	99

Computed Indices

	1t	1s	5s	1m	5m	15m
RV	28	30	33	33	30	47
RVac1	28	30	37	38	28	37
RK	32	31	31	24	22	12
MSRV	44	34	30	32	37	0
TSRV	28	24	20	37	52	8
MLRV	40	30	46	36	41	0
RRV	28	33	40	43	36	22
BR	41					
BPV	40	-	-	26	48	56
minRV	20	-	-	12	42	44
medRV	28	-	0	14	56	48
QRV	32	0	10	26	82	34
TrunRV	32	0	14	37	68	60

Note: This table aggregates, for groups of assets (either all 31 assets, or assets belonging to one class), the 90% Model Confidence Sets identifying the subset containing "best" estimators. Each table cell reports the percentage of all estimators of a certain estimator class and sampling frequency (across assets, and aggregating estimators using different sampling schemes and sampled price series) that are found to be in a Model Confidence Set. '-' indicates that for the assets under consideration, all estimators of that class and sampling frequency yield values that are unrealistically small and thus dropped from the competing set (see section 7.2 in the Appendix).

7 Appendix

7.1 Data cleaning

All series were cleaned according to a set of baseline rules similar to those in Barndorff-Nielsen et al. (2009). Using notation from that paper, these rules are:

- P1 Prices out of normal business hours were discarded.
- P2 Prices with a 1-tick reversal greater than 15 times the median spread were removed.
- P3 Prices were aggregated using the median of all prices with that time stamp.
- Q1 Quotes with bid above offer were removed.
- Q2 Quotes with a spread greater than 15 times the daily median spread were removed
- QT1 The maximum price was determined as the minimum of the maximum offer and the maximum transaction price, plus 2 times the daily median standard deviation. The minimum price was determined as the maximum of the minimum bid and the minimum transaction price, minus 2 times the daily spread. Transactions with prices outside of this range, or quotes where either price was outside this range were removed.
- QT2 Transactions with prices which were outside of the bid and offer over the previous 1 minute or subsequent 1 minute were removed. No action was taken if there were no quotes during this period.
- QT3 Quotes with bids above or offers below the observed trading price range over the previous and subsequent minute were removed.
- F1 The active future was chosen according to the highest transaction volume on each trading day, with the condition that once a future has been selected, it cannot be deselected in favor of a new contract and then reselected. When this occurred, the unique roll date was selected by maximizing the total transaction volume to choose a single roll date.

On the rare occasion that a problem was detected, the problematic data points were removed manually. Manual cleaning was needed in less than 0.1% of all days.

7.2 Additional summary statistics and results

This section summarizes some further summary statistics for the realized measures.

Our broad implementation of realized measures means that some questionable estimators are included, and for some of these measures, we see unrealistic estimates of QV (negative or zero values, for example) for several days. We use the following simple rule to remove the worst estimators before proceeding to formal rankings and tests: if values of the realized measure are less than a prespecified cutoff (0.0001 for interest rate and currency futures or 0.001 for all other assets) for more than 5% of the sample then that estimator is removed from the competing set, and not included in any subsequent analysis. Only 12 of the 31 assets had any realized measures removed, and the maximum number of removed measures was seven (out of 356 measures in total). Realized measures with a small number of unrealistic estimates are retained, and the values below the cutoff are replaced with the previous day's value. Table A2 records the estimators that are removed from each competing set for each asset according to this rule. Not surprisingly, these estimators include many that were implemented on an inappropriate sampling frequency relative to the frequency of the available price data.

Tables A3 and A4 supplement Table 2, providing summary statistics for each individual asset.

Table A5 presents information on the correlation between the estimators. As one would expect, the majority of the remaining estimators are highly correlated. On average, about half of the correlations are over 0.9, and about 25% are 0.95 or higher.

Table A6 presents correlation matrices for the ranks of individual realized measures, according to estimated accuracy, across pairs of assets in a given asset class. These rank correlations provide insights into whether the relative performance of realized measures is similar across assets in the same asset class.

Table A7 presents the size of the estimated model confidence set (MCS) for each individual asset.

Table A1
Short-hand codes for estimators.

Order:	Class _ SamplingFreq _ SamplingScheme PriceSeries _ Subsampling
Classes of Realized Measures	
RV	Realized Variance
BR	Realized Variance with Bandi-Russell Optimal Sampling
TSRV	Two-scales realized variance
MSRV	Multi-scales realized variance
RVac1	First-order autocorrelation adjusted realized variance
RKbart	Realized Kernel with flat-top Bartlett kernel
RKcub	Realized Kernel with flat-top cubic kernel
RKth2	Realized Kernel with flat-top Tukey-Hanning2 kernel
RKnfp	Realized Kernel with non-flat-top Parzen kernel
MLRV	Maximum-Likelihood realized variance
RRVm5	Realized range-based variance with block length 5
RRVm10	Realized range-based variance with block length 10
Sampling Frequency	
1t	tick-by-tick
1s	1-second
5s	5-second
1m	1-minute
5m	5-minute
15m	15-minute
Sampling Scheme	
c	calendar-time sampling
b	tick (business)-time sampling
Price series	
t	transactions prices
q	midquote
Subsampling	
ss	subsampled
"blank"	not subsampled
Example:	
RV_1m_ct_ss	Realized variance, using 1-minute calendar time sampling of trade prices, sub-sampled

Table A2

Non-jump robust estimators that were not implemented due to having a large number of very small or negative values

KO	MSRV_15m_ct	MSRV_15m_cq					
LSI	MSRV_15m_ct	MSRV_15m_cq	MSRV_15m_bq				
MSFT	MSRV_15m_ct	MSRV_15m_cq					
IFF	MSRV_15m_ct	MSRV_15m_bt	MSRV_15m_cq	MSRV_15m_bq	RKth2_1s_ct	RKth2_15m_bt	RKth2_15m_cq
SYY	MSRV_15m_ct	MSRV_15m_bt	MSRV_15m_cq	MSRV_15m_bq	RKth2_1s_ct	RKth2_15m_bt	
VOD	BRbc_cq						
SDR	TSRV_15m_bt	MSRV_15m_ct	MSRV_15m_bt	RKth2_1s_ct			
RSA	MSRV_15m_ct						
TU	MSRV_15m_cq						
US	RVac1_1t_bt						
JNI	BRbc_cq	BRbc_bq	RKbart_5s_bq	RKbart_1m_bq			
N225	MSRV_15m_c	MSRV_15m_b					

Table A3
Summary of Sample Means and Standard Deviations of Realized Measures

	Sample Mean				**Sample Standard Deviation**			
	median	std dev	min	max	median	std dev	min	max
KO	1.78	1.88	0.32	17.24	2.83	2.55	0.96	23.00
LSI	11.53	11.13	1.07	106.98	14.66	12.37	3.12	115.60
MSFT	3.09	3.91	0.84	31.34	4.18	5.32	1.59	37.90
IFF	2.78	2.29	0.24	21.72	5.18	3.49	0.56	30.66
SYY	2.38	2.83	0.41	24.49	3.51	4.44	0.65	37.56
DGE	2.73	3.49	0.60	31.37	4.11	5.46	1.43	42.47
VOD	4.44	9.51	1.38	68.50	5.71	10.53	2.28	70.50
SAB	3.60	2.62	1.21	22.40	6.45	4.46	1.83	29.24
SDR	8.92	6.23	0.89	47.60	17.29	10.54	1.67	70.39
RSA	6.96	7.09	1.03	61.13	11.99	9.36	2.71	81.44
TU	0.01	0.01	0.00	0.09	0.02	0.01	0.01	0.11
FV	0.06	0.06	0.03	0.53	0.08	0.05	0.03	0.56
TY	0.13	0.15	0.06	1.18	0.17	0.13	0.06	1.20
US	0.32	0.41	0.14	3.22	0.39	0.29	0.13	2.73
FGBL	0.09	0.12	0.03	0.95	0.08	0.09	0.03	0.73
FGBS	0.01	0.02	0.00	0.09	0.02	0.73	0.01	2.47
CD	0.33	0.39	0.15	3.17	0.35	0.39	0.16	3.12
AD	0.50	0.58	0.19	4.74	0.99	1.10	0.39	8.72
BP	0.23	0.27	0.10	2.15	0.30	0.35	0.14	2.72
URO	0.24	0.28	0.11	2.30	0.25	0.29	0.10	2.25
JY	0.28	0.31	0.12	2.56	0.36	0.35	0.16	2.88
STXE	1.77	2.13	0.57	16.58	3.20	4.09	1.38	29.25
JNI	1.16	1.79	0.42	14.03	1.73	1.90	0.55	15.03
FDX	1.75	2.03	0.63	16.54	2.80	3.02	1.31	23.97
FFI	1.29	1.57	0.61	12.31	2.09	2.54	0.96	19.06
ES	1.26	1.99	0.58	14.64	2.64	3.56	1.17	24.14
SPX	1.10	1.17	0.04	10.49	2.51	2.23	0.10	20.52
STOXX50E	1.50	1.52	0.08	13.63	2.53	2.29	0.17	20.02
DAX	1.78	1.88	0.35	16.45	3.03	2.88	1.04	24.47
FTSE	1.04	1.04	0.05	9.46	2.10	1.60	0.12	14.45
N225	0.91	0.68	0.02	7.67	1.41	1.18	0.04	12.51

Notes: The sample mean and standard deviation of each of the 398 (or 199) realized measures for all 31 assets were calculated. This table summarizes the summary statistics by listing the median sample mean, the standard deviation of the sample means, and the minimum and maximum values of sample means for a given asset. We do the same for the collection of 398 (or 199) sample standard deviations for each asset.

Table A4
Estimated autocorrelation of realized measures and quadratic variation

	rho(1)			rho(2)			rho*(1)			rho*(2)		
	mean	std dev	RV5m	mean	std dev	RV5m	mean	std dev	RV5m	mean	std dev	RV5m
KO	0.61	0.11	0.62	0.61	0.10	0.61	0.93	0.03	0.95	0.90	0.03	0.94
LSI	0.60	0.10	0.64	0.53	0.11	0.60	0.94	0.07	0.98	0.89	0.12	0.96
MSFT	0.73	0.11	0.72	0.71	0.11	0.68	0.96	0.02	0.96	0.94	0.02	0.95
IFF	0.49	0.14	0.46	0.45	0.15	0.41	0.95	0.01	0.93	0.92	0.02	0.93
SYY	0.56	0.07	0.57	0.53	0.11	0.53	0.91	0.03	0.91	0.88	0.04	0.90
DGE	0.60	0.11	0.61	0.54	0.11	0.49	0.97	0.02	0.98	0.95	0.02	0.97
VOD	0.67	0.11	0.45	0.60	0.12	0.44	0.97	0.01	0.96	0.96	0.02	0.95
SAB	0.50	0.12	0.49	0.41	0.12	0.33	0.96	0.03	0.97	0.94	0.04	0.91
SDR	0.48	0.11	0.59	0.38	0.10	0.48	0.93	0.03	0.95	0.91	0.04	0.94
RSA	0.57	0.11	0.56	0.53	0.10	0.50	0.97	0.02	0.96	0.95	0.01	0.93
TU	0.37	0.16	0.35	0.36	0.14	0.35	0.96	0.02	0.94	0.95	0.02	0.95
FV	0.24	0.15	0.20	0.23	0.13	0.17	0.96	0.02	0.95	0.94	0.02	0.94
TY	0.30	0.18	0.19	0.27	0.16	0.16	0.97	0.01	0.96	0.94	0.02	0.94
US	0.28	0.17	0.17	0.24	0.15	0.13	0.96	0.02	0.94	0.93	0.03	0.92
FGBL	0.55	0.15	0.60	0.48	0.13	0.52	0.97	0.01	0.96	0.93	0.01	0.91
FGBS	0.29	0.30	0.58	0.25	0.26	0.49	0.93	0.16	0.96	0.82	0.28	0.94
CD	0.70	0.11	0.68	0.68	0.10	0.68	1.00	0.01	1.00	0.98	0.01	0.97
AD	0.74	0.07	0.71	0.76	0.05	0.78	0.93	0.05	0.94	0.90	0.04	0.93
BP	0.75	0.10	0.71	0.72	0.08	0.70	0.99	0.01	0.99	0.98	0.01	0.98
URO	0.65	0.12	0.63	0.60	0.12	0.58	0.98	0.01	0.98	0.96	0.01	0.95
JY	0.55	0.13	0.50	0.44	0.12	0.40	0.95	0.01	0.95	0.91	0.02	0.93
STXE	0.45	0.26	0.61	0.40	0.23	0.54	0.95	0.02	0.95	0.94	0.02	0.94
JNI	0.70	0.11	0.70	0.66	0.10	0.63	0.91	0.05	0.86	0.93	0.04	0.87
FDX	0.64	0.15	0.23	0.58	0.15	0.20	0.95	0.01	0.96	0.95	0.02	0.95
FFI	0.73	0.08	0.71	0.69	0.07	0.65	0.96	0.01	0.97	0.94	0.01	0.94
ES	0.68	0.09	0.68	0.67	0.08	0.67	0.90	0.03	0.87	0.86	0.04	0.85
SPX	0.66	0.08	0.69	0.65	0.08	0.68	0.91	0.03	0.92	0.86	0.03	0.86
STOXX50E	0.67	0.09	0.57	0.64	0.07	0.57	0.94	0.03	0.90	0.93	0.03	0.89
DAX	0.69	0.08	0.70	0.60	0.08	0.62	0.94	0.01	0.96	0.94	0.02	0.96
FTSE	0.53	0.10	0.55	0.53	0.09	0.51	0.90	0.05	0.89	0.87	0.06	0.85
N225	0.70	0.10	0.74	0.65	0.08	0.67	0.95	0.04	0.95	0.94	0.03	0.94
Average	0.57	0.12	0.56	0.53	0.12	0.51	0.95	0.03	0.95	0.92	0.04	0.93

Notes: This table lists the mean and standard deviation, by asset, of sample autocorrelations of realized measures (denoted "rho") and the estimated autocorrelation of QV based on a realized measure (denoted "rho*"), using the instrumental variables method of Hansen and Lunde (2010). The estimates based purely on RV5min are also presented.

Table A5
Quantiles of pairwise correlations between realized measures of a given asset

	0.01	**0.05**	**0.10**	**0.25**	**0.50**	**0.75**	**0.90**	**0.95**	**0.99**
KO	0.53	0.63	0.68	0.76	0.85	0.92	0.96	0.98	0.99
LSI	0.37	0.48	0.57	0.70	0.82	0.91	0.96	0.97	0.99
MSFT	0.30	0.50	0.62	0.76	0.87	0.94	0.98	0.99	1.00
IFF	0.45	0.52	0.60	0.70	0.86	0.96	0.98	0.99	1.00
SYY	0.47	0.62	0.67	0.79	0.89	0.94	0.97	0.98	1.00
DGE	0.48	0.60	0.65	0.72	0.81	0.90	0.94	0.96	0.99
VOD	0.38	0.65	0.70	0.77	0.87	0.92	0.96	0.97	0.99
SAB	0.22	0.38	0.48	0.62	0.73	0.82	0.89	0.93	0.99
SDR	0.21	0.37	0.47	0.61	0.72	0.81	0.90	0.95	1.00
RSA	0.54	0.64	0.70	0.78	0.84	0.90	0.94	0.96	0.99
TU	0.49	0.58	0.63	0.72	0.81	0.89	0.94	0.96	0.99
FV	0.39	0.51	0.56	0.66	0.76	0.86	0.92	0.95	0.98
TY	0.41	0.52	0.58	0.71	0.83	0.91	0.96	0.97	0.99
US	0.30	0.43	0.51	0.65	0.80	0.91	0.96	0.97	0.99
FGBL	0.45	0.57	0.62	0.72	0.83	0.91	0.95	0.97	0.99
FGBS	0.00	0.02	0.02	0.05	0.64	0.96	1.00	1.00	1.00
CD	0.63	0.79	0.83	0.88	0.93	0.96	0.98	0.99	1.00
AD	0.73	0.85	0.88	0.92	0.96	0.98	0.99	0.99	1.00
BP	0.75	0.83	0.86	0.91	0.94	0.97	0.99	0.99	1.00
URO	0.66	0.74	0.79	0.86	0.92	0.96	0.98	0.99	1.00
JY	0.68	0.77	0.81	0.87	0.92	0.96	0.98	0.99	1.00
STXE	0.16	0.22	0.28	0.46	0.78	0.94	0.98	0.99	1.00
JNI	0.38	0.62	0.70	0.81	0.88	0.93	0.96	0.98	0.99
FDX	0.45	0.57	0.66	0.79	0.90	0.95	0.98	0.99	1.00
FFI	0.81	0.88	0.90	0.93	0.96	0.98	0.99	0.99	1.00
ES	0.74	0.84	0.87	0.92	0.96	0.98	0.99	1.00	1.00
SPX	0.68	0.81	0.84	0.89	0.93	0.97	0.99	0.99	1.00
STOXX50E	0.65	0.76	0.81	0.87	0.92	0.96	0.98	0.99	1.00
DAX	0.57	0.72	0.79	0.87	0.92	0.96	0.98	0.99	1.00
FTSE	0.46	0.64	0.72	0.80	0.88	0.95	0.98	0.99	1.00
N225	0.59	0.69	0.75	0.86	0.93	0.97	0.99	1.00	1.00
Average	0.48	0.60	0.66	0.75	0.86	0.93	0.97	0.98	1.00

Note: All values of "1.00" are due to rounding. Actual correlation values are less than 1.

Table A6
Cross-asset correlations of rankings

Individual Equities

	KO	**LSI**	**MSFT**	**IFF**	**SYY**	**DGE**	**VOD**	**SAB**	**SDR**	**RSA**
KO	1	0.91	0.87	0.76	0.86	0.64	0.68	0.78	0.67	0.67
LSI		1	0.84	0.88	0.94	0.48	0.50	0.77	0.63	0.52
MSFT			1	0.77	0.78	0.58	0.65	0.69	0.59	0.59
IFF				1	0.87	0.32	0.38	0.61	0.49	0.40
SYY					1	0.42	0.40	0.73	0.65	0.47
DGE						1	0.88	0.76	0.78	0.95
VOD							1	0.65	0.61	0.87
SAB								1	0.90	0.77
SDR									1	0.84
RSA										1

Avg Corr: 0.68

Currency Futures

	CD	**AD**	**BP**	**URO**	**JY**
CD	1	0.94	0.83	0.83	0.91
AD		1	0.83	0.76	0.85
BP			1	0.87	0.87
URO				1	0.95
JY					1

Avg Corr: 0.87

Index Futures

	STXE	**JNI**	**FDX**	**FFI**	**ES**
STXE	1	0.88	0.63	0.64	0.89
JNI		1	0.51	0.49	0.76
FDX			1	0.96	0.80
FFI				1	0.79
ES					1

Avg Corr. 0.74

Bond Futures

	TU	**FV**	**TY**	**US**	**FGBL**	**FGBS**
TU	1	0.96	0.89	0.86	0.83	0.69
FV		1	0.96	0.90	0.88	0.63
TY			1	0.91	0.93	0.59
US				1	0.95	0.79
FGBL					1	0.70
FGBS						1

Avg Corr: 0.83

Computed Indices

	SPX	**S50E**	**DAX**	**FTSE**	**N225**
SPX	1	0.97	0.80	0.94	0.73
S50E		1	0.80	0.96	0.75
DAX			1	0.84	0.84
FTSE				1	0.70
N225					1

Avg Corr: 0.83

Table A7
Size of 90% Model Confidence Sets (QLIKE loss)

		Proxy for QV									
		DailyRV		*15min RV*		*5min RV*		*1min MSRV*		*1min RKth2*	
Asset	Total # Estimators	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
KO	396	114	28.8	8	2.0	11	2.8	8	2.0	8	2.0
LSI	395	30	7.6	8	2.0	10	2.5	6	1.5	8	2.0
MSFT	396	75	18.9	6	1.5	3	0.8	5	1.3	5	1.3
IFF	391	125	32.0	32	8.2	27	6.9	49	12.5	49	12.5
SYY	392	143	36.5	14	3.6	20	5.1	8	2.0	10	2.6
DGE	398	28	7.0	13	3.3	2	0.5	14	3.5	13	3.3
VOD	397	35	8.8	5	1.3	5	1.3	4	1.0	4	1.0
SAB	398	80	20.1	15	3.8	13	3.3	5	1.3	5	1.3
SDR	394	21	5.3	3	0.8	4	1.0	2	0.5	3	0.8
RSA	397	21	5.3	5	1.3	9	2.3	16	4.0	8	2.0
TU	397	10	2.5	27	6.8	16	4.0	30	7.6	31	7.8
FV	398	38	9.5	5	1.3	6	1.5	4	1.0	4	1.0
TY	398	20	5.0	20	5.0	24	6.0	24	6.0	23	5.8
US	397	8	2.0	15	3.8	6	1.5	15	3.8	9	2.3
FGBL	398	3	0.8	5	1.3	4	1.0	13	3.3	11	2.8
FGBS	398	37	9.3	10	2.5	2	0.5	18	4.5	16	4.0
CD	398	33	8.3	4	1.0	6	1.5	6	1.5	4	1.0
AD	398	133	33.4	6	1.5	5	1.3	5	1.3	8	2.0
BP	398	13	3.3	20	5.0	23	5.8	9	2.3	10	2.5
URO	398	9	2.3	13	3.3	6	1.5	6	1.5	6	1.5
JY	398	18	4.5	16	4.0	11	2.8	16	4.0	16	4.0
STXE	398	16	4.0	12	3.0	10	2.5	4	1.0	10	2.5
JNI	394	14	3.6	1	0.3	2	0.5	1	0.3	1	0.3
FDX	398	18	4.5	10	2.5	11	2.8	5	1.3	5	1.3
FFI	398	19	4.8	7	1.8	17	4.3	17	4.3	20	5.0
ES	398	34	8.5	8	2.0	13	3.3	9	2.3	1	0.3
SPX	398	4	1.0	4	1.0	4	1.0	22	5.5	4	1.0
STOXX50E	398	30	7.5	10	2.5	2	0.5	16	4.0	18	4.5
DAX	398	40	10.1	2	0.5	4	1.0	22	5.5	14	3.5
FTSE	398	34	8.5	24	6.0	26	6.5	22	5.5	20	5.0
N225	394	44	11.2	12	3.0	4	1.0	10	2.5	4	1.0

Notes: Columns (a) display the number of estimators included in a MCS, and columns (b) display the percentage of total estimators that are included in a MCS.

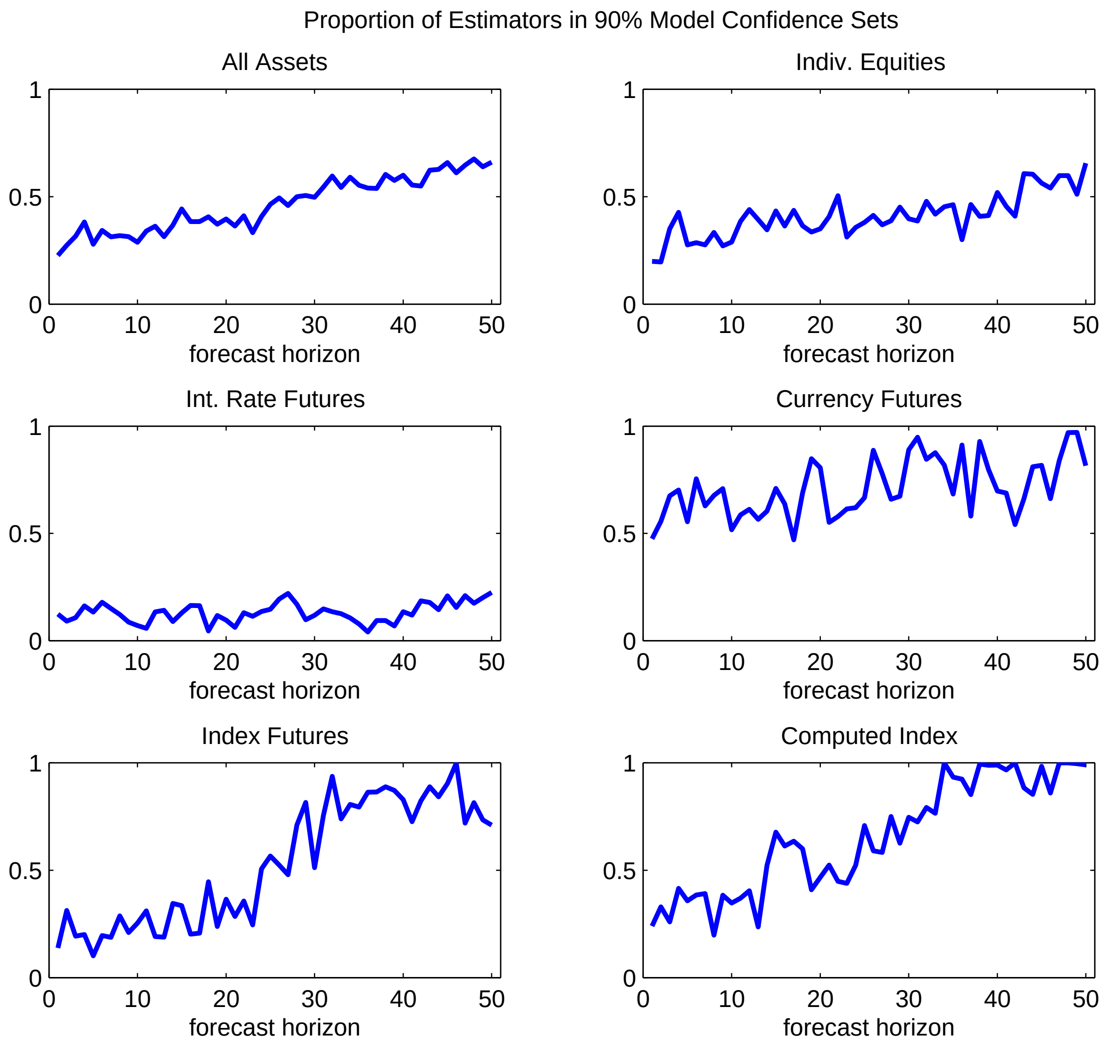

Figure 1: *This figure presents the proportion of all 604 realized measures included in the 90% model confidence set at each forecast horizon, ranging from 1 to 50 days. The upper left panel presents the results across all 31 assets, and the remaining panels present results for each of the 5 asset classes separately.*

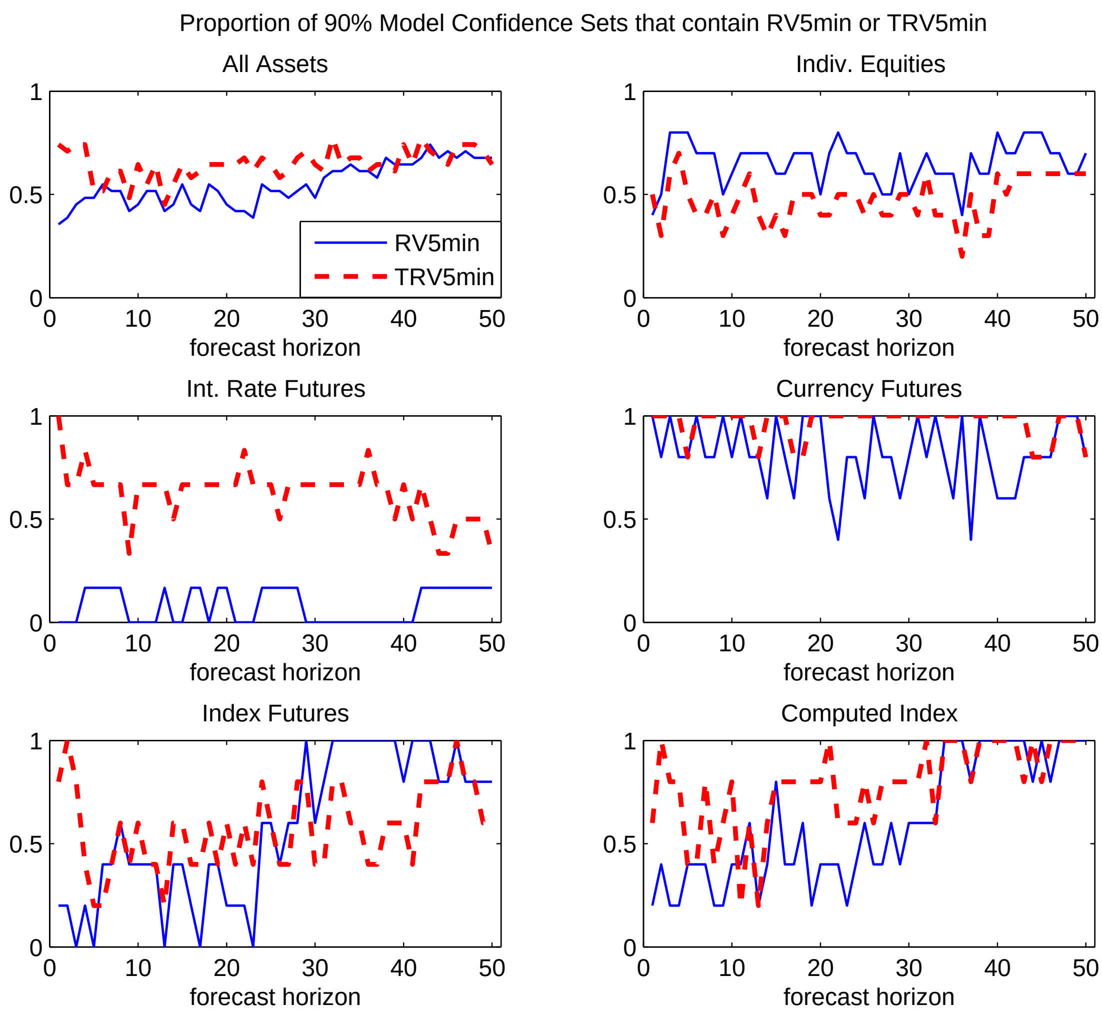

Figure 2: *This figure presents the proportion of 90% model confidence sets (across assets) that contain 5-minute RV and 5-minute truncated RV, at each forecast horizon ranging from 1 to 50 days. The upper left panel presents the results across all 31 assets, and the remaining panels present results for each of the 5 asset classes separately.*