PREDICTING VOLATILITY:
GETTING THE MOST OUT OF RETURN DATA
SAMPLED AT DIFFERENT FREQUENCIES

Eric Ghysels
Pedro Santa-Clara
Rossen Valkanov

Predicting Volatility: Getting the Mos out of Return Data Sampled at Different Frequencies
Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov
NBER Working Paper No. 10914
November 2004
JEL No. G1

ABSTRACT

We consider various MIDAS (Mixed Data Sampling) regression models to predict volatility. The
models differ in the specification of regressors (squared returns, absolute returns, realized volatility,
realized power, and return ranges), in the use of daily or intra-daily (5-minute) data, and in the length
of the past history included in the forecasts. The MIDAS framework allows us to compare models
across all these dimensions in a very tightly parameterized fashion. Using equity return data, we find
that daily realized power (involving 5-minute absolute returns) is the best predictor of future
volatility (measured by increments in quadratic variation) and outperforms model based on realized
volatility (i.e. past increments in quadratic variation). Surprisingly, the direct use of high-frequency
(5-minute) data does not improve volatility predictions. Finally, daily lags of one to two months are
sucient to capture the persistence in volatility. These findings hold both in- and out-of-sample.

Eric Ghysels
University of North Carolina, Chapel Hill
Gardner Hall, CB 3305
Chapel Hill, NC 27599-3305
eghysels@unc.edu

Pedro Santa-Clara
Anderson School
UCLA
110 Westwood Plaza, Suite C4.21
Los Angeles, CA 90095-1481
and NBER
pedro.santa-clara@anderson.ucla.edu

Rossen Valkanov
Anderson School
UCLA
110 Westwood Plaza, Suite C4.21
Los Angeles, CA 90095-1481
rossen.valkanov@anderson.ucla.edu

The conditional volatility literature, starting with Engle's (1982) ARCH-class of models, has been successful at capturing the dynamics of return variance using simple parametric models. A measure of that success is the widespread use of such models in all areas of finance by academics and practitioners alike. And while most researchers would agree that it is important to have a good prediction model of conditional volatility, the question of what model to use is still unsettled.

When it comes to forecasting volatility, there are many existing models in addition to the benchmark ARCH/GARCH models of Engle (1982) and Bollerslev (1986) which cast future variance as a polynomial of past squared returns, i.e., $\hat{\sigma}^2_{t+1|t} \equiv A(L)r^2_t$. One alternative is to look for variables, other than squared returns, that relate to future volatility. Ding et al. (1993) and several others show that low-frequency components of volatility might be better captured by absolute returns instead of squared returns. Also, Alizadeh et al. (2002) and Gallant et al. (1999) find daily ranges (high-low price ranges) to be good predictors of volatility. Another rapidly growing research area focuses on data-driven models of realized volatility computed from intra-daily returns sampled at very short intervals such as 5 minutes (Andersen and Bollerslev (1998)).[1] All these models suggest a variety of possible ways to forecast volatility. Hence, it seems natural to ask whether some of the suggested predictors are clearly dominated by others and whether there are real benefits from using high-frequency data.[2] These questions have proven difficult to answer because the models considered are so different in terms of regressors, frequencies, parameterizations, and return histories, that is it difficult to directly compare them.

We use Mixed Data Sampling (henceforth MIDAS) regression models introduced in Ghysels, Santa-Clara and Valkanov (2002a,b) to provide answers to these questions. MIDAS regressions allow us to run parsimoniously parameterized regressions of data observed at different frequencies. There are several advantages of using mixed data sampling regressions. They allow us to study, in a unified framework, the forecasting performance of a large class of volatility models which involve: (i) data sampled at different frequencies; (ii) various past data window lengths; and (iii) different regressors. The specification of the regressions

---

[1] Also, see Andersen et al. (2001, 2002, 2003), Andreou and Ghysels (2002), Barndorff-Nielsen and Shephard (2001, 2002a,b, 2003a), Taylor and Xu (1997), among others.

[2] High-frequency data also suffer from microstructure artifacts, such as bid-ask bounce (Roll, 1984), screen fighting, jumps, and irregular or missing data, all of which can lead to biases in the volatility estimates. The impact of microstructure noise on realized variance has been recently studied by Aït-Sahalia, Mykland and Zhang (2003), Hansen and Lunde (2004), Bandi and Russell (2003), Zhang, Mykland and Aït-Sahalia (2003), among others.

combine recent developments regarding estimation of volatility and a not so recent literature on distributed lag models.[3] We focus on predicting future conditional variance, measured as increments in quadratic variation (or its log transformation) from one week to one month horizons, because these are the horizons that are most widely used for option pricing, portfolio management, and hedging applications.

First, we use MIDAS regressions to examine whether future volatility is well predicted by past daily squared returns, absolute daily returns, realized daily volatility, realized daily power (sum of intra-daily absolute returns, a measure proposed by Barndorff-Nielsen and Shephard (2003b, 2004) and Woerner (2002)), and daily range. Since all of the regressors are used within a framework with the same number of parameters and the same maximum number of lags, the results from MIDAS regressions are directly comparable. Hence, the MIDAS setup allows us to determine if one of the regressors dominates others. We find that, for the Dow Jones Index and six individual stock return series, the realized power clearly dominates all other daily predictors of volatility at all horizons. Importantly, the predictive content of the realized power is evident not only from in-sample goodness of fit measures, but also from out-of-sample forecasts. The daily range is also a good predictor in the sense that it dominates squared and absolute daily returns. Our method is a significant departure from the usual autoregressive model building approach embedded in the ARCH literature and its recent extensions such as high-frequency data-based approaches. A comparison of the MIDAS regressions with purely autoregressive volatility models reveals that the MIDAS forecasts are better at forecasting future realized volatility in- and out-of-sample.

Second, the weights in the MIDAS regressions are parameterized by a flexible function. Obviously, the choice of regressors is as important as is the profile of weights placed on them. In our MIDAS framework, the shape of the weight function is determined by two parameters that are estimated from the data. Hence, the weight profile on the lagged predictors is captured by the shape of the function, whose parameters are estimated from the data with no additional pre-testing or lag-selection procedures. We find that daily lags longer than about 50 days do not help (nor hurt) the forecasts, for any of the regressors.

Third, mixed data regressions allow us to directly project future realized volatility onto high-frequency (say 5-minute) squared and absolute returns without daily pre-filtering and

---

[3]See e.g. Dhrymes (1971) and Sims (1974) for surveys on distributed lag models. Many econometrics textbooks also cover the topic, see e.g. Greene (2000, chap. 17), Judge et al. (1985, chap. 9 - 10), Stock and Watson (2003, chap. 13) Wooldridge (2000, chap. 18), among others.

without increasing the number of parameters. Hence, we are able to analyze if there are real benefits from directly using high-frequency data in volatility forecasting. Surprisingly, we find that forecasts using high-frequency data directly do not outperform those that use daily regressors (although the daily regressors are themselves obtained through the aggregation of high-frequency data). It must be noted that none of these results are driven by over-fitting or parameter proliferation. Indeed, all MIDAS specifications – daily and 5-minutes – are directly comparable since they all have the same number of estimated parameters.

In summary, we find that daily realized power is the best predictor of future increments in quadratic variation, followed by the daily range. The prediction equations involve about 50 daily lags and there is no real benefit of using intra-daily data directly. These MIDAS regressions also outperform other linear forecast models involving daily realized volatility. Finally, all of the above results hold in- and out-of-sample. The out-of-sample forecasting precision, which is perhaps the ultimate measure of a model's forecasting potential, indicates that our results are unlikely to be due to sampling error or over-fitting.

The MIDAS regressions can also be used to model asymmetries and the joint forecasting power of the regressors. In fact, Engle and Gallo (2003) use the multiplicative error model (MEM) of Engle (2002) and find improvements in forecasting volatility from the joint use of absolute returns, daily ranges, and realized volatilities using S&P 500 index returns data. Interestingly enough, their results agree with ours, despite the different data set and different method, as they argue that range-based measures in particular provide a very good forecast of future volatility.

The paper is structured as follows. In a first section we introduce and discuss MIDAS volatility models. In section two, we use daily regressors to forecast weekly to monthly volatility in the MIDAS framework. The third section is devoted to MIDAS volatility forecasts involving intra-daily data. Section four concludes.

1   MIDAS Models of Conditional Volatility

To fix notation, let daily returns be denoted by $r_{t,t-1} = \log(P_t) - \log(P_{t-1})$. Throughout the paper the time index t will refer to daily sampling. When the data is sampled at a higher frequency, say, m-times in a day, we will denote the return over this interval as $r_{t,t-1/m} = \log(P_t) - \log(P_{t-1/m})$. For instance, in our study, returns are sampled every five minutes

between the trading hours of 9:30 am and 4:05 pm (corresponding to 80 five-minute intervals within a trading day), and we will write $r_{t,t-1/80} = \log(P_t) - \log(P_{t-1/80})$, which corresponds to the last 5-minute return of day $t - 1$.

Our goal is to predict a measure of volatility over some future horizon H, $V_{t+H,t}$. As a primary measure of volatility for the period t to $t + H$, we consider the increments in the quadratic variation of the return process $Q_{t+H,t}$. We focus on predicting future realized volatility from one week ($H = 5$) to one month ($H = 20$) horizon. These are horizons that matter mostly for option pricing and portfolio management. Focusing on predicting future increments of quadratic variation also allows us to make our analysis directly comparable with a large body of existing literature. The quadratic variation is not observed directly but can be measured with some discretization error. One such measure would be the sum of (future) squared returns, namely $\sum_{j=1}^{Hm}[r_{(t+H)-(j-1)/m,(t+H)-(j-2)/m}]^2$, which we will denote by $\tilde{Q}_{t+H,t}^{(Hm)}$ since it involves a discretization based on Hm intra-daily returns. The superscript in parentheses indicates the number of high-frequency data used to compute the variable.

Besides increments in quadratic variation, we also consider $\log\left(\tilde{Q}_{t+H,t}^{(Hm)}\right)$ as a target variable to forecast. Previous papers, including Andersen et al. (2003), have observed that forecasting the log transformation yields better in- and out-of-sample forecasts of the variance as it puts less weight on extreme realizations of the quadratic variation.[4]

1.1 The MIDAS Specification

A daily MIDAS volatility model is a regression model:

$$V_{t+H,t}^{(Hm)} = \mu_H + \phi_H \sum_{k=0}^{k^{max}} b_H(k,\theta)\tilde{X}_{t-k,t-k-1}^{(m)} + \varepsilon_{Ht} \tag{1.1}$$

where $V_{t+H,t}^{(Hm)}$ is a measure of (future) volatility such as $\tilde{Q}_{t+H,t}^{(Hm)}$ or $\log\left(\tilde{Q}_{t+H,t}^{(Hm)}\right)$. Specification (1.1) has three important features when compared to other models of conditional volatility (discussed below). First, the volatility measure on the left-hand side, $V_{t+H,t}^{(Hm)}$, and the variables on the right-hand side, $\tilde{X}_{t-k,t-k-1}^{(m)}$, might be sampled at different frequencies. Second, the polynomial lag parameters b_H are parameterized to be a function of θ, thereby

---

[4]We also considered MIDAS regressions with $(\tilde{Q}_{t+H,t}^{(Hm)})^{1/2}$ as left hand side variable. The results are omitted here to save space. The conclusions concur with those reported in the empirical section below.

allowing for a longer history without a proliferation of parameters. Third, MIDAS regressions typically do *not* exploit an autoregressive scheme, so that $\tilde{X}_{t-k,t-k-1}^{(m)}$ is not necessarily related to lags of the left hand side variable. Instead, MIDAS regressions are first and foremost regression models and therefore the selection of $\tilde{X}_{t-k,t-k-1}^{(m)}$ amounts to choosing the best predictor of future quadratic variation from the set of several possible measures of past fluctuations in returns. MIDAS regressions could potentially involve more than one type of regressors, see Ghysels et al. (2004) for further discussion. MIDAS regression models may also be nonlinear and indeed some of the regressors we will consider may provide better results with nonlinear specifications. For instance, Engle and Gallo (2003) provide interesting results along these lines. For simplicity, in this paper we consider only a single regressor linear MIDAS setting.

Sampling at Different Frequencies

In equation (1.1), the volatility is measured at weekly, bi-weekly, tri-weekly, and monthly frequency, whereas the forecasting variables $\tilde{X}_{t-k,t-k-1}^{(m)}$ are available at higher frequencies. For instance, we can use daily data to compute a forecast of next month's volatility ($H = 22$). In other words, the return volatility over the month of, say, April (from the close of the market during the last day of March to the close of the market during the last day of April) will be forecasted with daily data up to the last day of March. But we could also use hourly or five-minute data to form monthly volatility forecasts. Thus, our model allows us not only to forecast volatility with data sample at different frequencies, but also to compare such forecasts and ultimately to see whether Merton's (1980) well-known continuous asymptotic arguments hold up in practice.

In general, the MIDAS framework allows us to investigate whether the use of high-frequency data necessarily leads to better volatility forecasts at various horizons. These issues have motivated much of the recent literature on high-frequency data, see Andersen et al. (2001, 2002, 2003), Andreou and Ghysels (2002), Barndorff-Nielsen and Shephard (2001, 2002a,b, 2003a), among others. In some cases, the right-hand side variables are computed using m high-frequency returns, in which case they are denoted by a superscript (m). For instance, if we want to compute a monthly forecast of volatility using lagged daily volatility estimates obtained from five-minute data, m would be equal to 80, the number of five-minutes in a day.

The MIDAS volatility models allow for a great degree of flexibility. For the sake of systematizing the results, in the next section we consider forecasts at weekly, bi-weekly, tri-weekly, and monthly frequency using daily data. In a subsequent section, we will turn to intra-daily regressors.

Parsimony of Parameterization

Another distinguishing feature of (1.1) is that the lag coefficients $b_H(k, \theta)$ (weights) are *not* unrestricted parameters. Rather they are parameterized as a function of θ, where θ is a small-dimensional vector. A flexible parameterization is an important element in the MIDAS specification, as the inclusion of high-frequency data might imply a significant increase in the number of lagged forecasting variables and unrestricted parameters to estimate. It allows us to dramatically reduce the number of parameters to estimate, which is particularly relevant in estimating a persistent process, such as volatility, where distant $\tilde{X}^{(m)}_{t-k,t-k-1}$ are likely to have an impact on current volatility.

Even with daily forecasting variables, the unrestricted specification of the weights results in a lot of parameters to estimate. The problem only worsens with higher-frequency data. As we will see below, a suitable parameterization $b_H(k, \theta)$ circumvents the problem of parameter proliferation and of choosing the truncation point k^{max}. Hence, the parameterization $b_H(k, \theta)$ is one of the most important ingredients in a MIDAS regression.

The weights $b_H(k; \theta)$ are normalized to add up to one, which allows us to estimate a scale parameter ϕ_H. Another restriction that is very useful in the context of volatility models is to insure that the weights $b_H(k; \theta)$ be non-negative. Such a condition guarantees that the volatility process itself is non-negative.[5] In general, there are many ways of parameterizing $b_H(k, \theta)$. We focus on one particular specification based on the Beta function, which has only two parameters, or $\theta = [\theta_1; \theta_2]$:

$$b_H(k; \theta) = \frac{f(\frac{k}{k^{max}}, \theta_1; \theta_2)}{\sum_{j=1}^{k^{max}} f(\frac{j}{k^{max}}, \theta_1; \theta_2)} \tag{1.2}$$

where: $f(z, a, b) = z^{a-1}(1 - z)^{b-1}/\beta(a, b)$ and $\beta(a, b)$ is based on the Gamma function, or

---

[5]It should parenthetically be noted that Nelson and Cao (1992) note that imposing non-negativity constraints on GARCH parameters is unnecessarily restrictive. They do argue, however, that non-negativity of the Wold representation weights is necessary. Restricting the weights $b_H(k; \theta)$ to be non-negative is very similar to the condition imposed by Nelson and Cao.

$\beta(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. Specification (1.2) was introduced in Ghysels et al. (2002b, 2004) and has several important characteristics. Namely, (1) it provides positive coefficients, which is necessary for *a.s.* positive definiteness of the estimated volatility, (2) with $\theta_1 = 1$ and $\theta_2 > 1$ one has a slowly decaying pattern typical of volatility filters, which means that only one parameter is left to determine the shape, and (3) with $\theta_1 = \theta_2 = 1$ one has equal weights, which corresponds to a rolling estimator of the volatility. The flexibility of the Beta function is well known. It is often used in Bayesian econometrics to impose flexible, yet parsimonious prior distributions. The function can take many shapes, including flat weights, gradually declining weights as well as hump-shaped patterns. While MIDAS regression models are not limited to Beta distributed lag schemes, for our purpose we focus our attention on this specification. We refer to Ghysels et al. (2002b, 2004) for alternative weight specifications and further details.

The parameterization also allows us to compare MIDAS models at different frequencies as the number of parameters to estimate will be the same even though the weights on the data and the forecasting capabilities might differ across horizons. We don't have to adjust our measures of fit for the number of parameters. In all estimations, we have either one or two parameters determining the pattern of the weights, the former being the case when we restrict our attention to $\theta_1 = 1$ and only estimate $\theta_2 > 1$.

To illustrate the issue of parameter proliferation, consider Figure 1. It displays the estimated *unconstrained* parameters of equation (1.1) for lags up to 10 days. The figure contains results from various regressors $\tilde{X}_{t-k,t-k-1}^{(m)}$, such as $\tilde{Q}_{t-k,t-k-1}^{(m)}$, as well as absolute daily returns, daily range, and daily realized power, all of which we discuss at length below. We notice from the results displayed in the figure that the parameter estimates appear to be erratic as the lag increases. Hence, volatility models such as (1.1), whose weights are *not* tightly parameterized, do only well even with a small number of lags and almost surely will produce poor out-of-sample forecasts. It must be noted that the robust performance of ARCH/GARCH models can largely be attributed to capturing the dynamics of a large number of past shocks with only a few parameters. This basic idea is also the insight behind the MIDAS regressions.

Various Regressors

In the MIDAS volatility model (1.1), $\tilde{X}_{t-k,t-k-1}^{(m)}$ can be any variable that has the ability to forecast $\tilde{Q}_{t+H,t}^{(Hm)}$. To put it differently, MIDAS volatility regressions can involve $\tilde{X}_{t-k,t-k-1}^{(m)}$

other than past squared returns or past realized volatility, which are the usual regressors considered in the autoregressive conditional volatility literature. The MIDAS approach puts us in the mind set of regression analysis and prompts us to explore various regressors that have the potential to predict future volatility. A number of predictors other than past realized volatility or squared returns have been proposed in various models. Unlike MIDAS, however, these models are typically autoregressive in nature. The MIDAS setup allows us to compare the forecasting ability of different $\tilde{X}_{t-k,t-k-1}^{(m)}$'s and to choose the model with the best forecasting ability.

In the context of forecasting the quadratic variation $\tilde{Q}_{t+H,t}^{(Hm)}$, we consider the following regressors

$$\tilde{Q}_{t+H,t}^{(Hm)} = \mu_H^Q + \phi_H^Q \sum_{k=0}^{k^{max}} b_H^Q(k,\theta)\tilde{Q}_{t-k,t-k-1}^{(m)} + \varepsilon_{Ht}^Q \qquad (1.3)$$

$$\tilde{Q}_{t+H,t}^{(Hm)} = \mu_H^G + \phi_H^G \sum_{k=0}^{k^{max}} b_H^G(k,\theta)(r_{t-k,t-k-1})^2 + \varepsilon_{Ht}^G \qquad (1.4)$$

$$\tilde{Q}_{t+H,t}^{(Hm)} = \mu_H^a + \phi_H^a \sum_{k=0}^{k^{max}} b_H^a(k,\theta)|r_{t-k,t-k-1}| + \varepsilon_{Ht}^a \qquad (1.5)$$

$$\tilde{Q}_{t+H,t}^{(Hm)} = \mu_H^r + \phi_H^r \sum_{k=0}^{k^{max}} b_H^r(k,\theta)[hi-lo]_{t-k,t-k-1} + \varepsilon_{Ht}^r \qquad (1.6)$$

$$\tilde{Q}_{t+H,t}^{(Hm)} = \mu_H^p + \phi_H^p \sum_{k=0}^{k^{max}} b_H^P(k,\theta)\tilde{P}_{t-k,t-k-1}^{(m)} + \varepsilon_{Ht}^p \qquad (1.7)$$

In equation (1.3), past $\tilde{Q}_{t,t-1}^{(m)}$ are used to predict $\tilde{Q}_{t+H,t}^{(Hm)}$. Examples of such models have been advocated by Andersen et al. (2001, 2002, 2003) and are discussed at length below. Specification (1.4) is a projection of $\tilde{Q}_{t+H,t}^{(Hm)}$ onto lagged daily returns and corresponds to the ARCH/GARCH class of models (under some parameter restrictions).[6]

Equations (1.5) and (1.6) involve projecting $\tilde{Q}_{t+H,t}^{(Hm)}$ onto past daily absolute returns and daily ranges, respectively, which are two alternative measures of volatility. Therefore they are natural candidate regressors in the MIDAS specification. It is often argued that in the presence of deviations from normality absolute values could be more robust than squared

---

[6]Technically speaking, projections involve regressions without a constant term. Here and in the remainder of the paper projections will include a constant as regressor.

values for conditional variance estimation (see e.g. Davidian and Caroll (1987)) whereas the virtues of daily range have been explored most recently by Alizadeh et al. (2001) and Gallant et al. (1999). Typically, past absolute returns (ranges) are used to predict future absolute returns (ranges). In particular, when absolute returns (daily ranges) are considered, the autoregressive features of absolute returns (daily ranges) are studied and modeled (see e.g. Ding et al. (1993) for absolute returns and Alizadeh et al. (2001) for daily ranges). Hence, the exploration of alternative measures of volatility has been cast in the context of autoregressive schemes. Here we introduce absolute returns and ranges as alternative predictors and examine their success (relative to the other predictors) at predicting realized volatility. The MIDAS regression format makes this a relatively straightforward exercise.

The last regression (1.7) involves similar arguments using techniques and developments of more recent date. The preference for absolute returns is a subject that has received much attention recently, see in particular Barndorff-Nielsen and Shephard (2003b, 2004) and Woerner (2002). Recall that $\tilde{Q}_{t,t+1}^{(m)}$ is defined as the sum of m intra-daily squared returns. Instead of taking squared returns, say every five minutes, Barndorff-Nielsen and Shephard suggest to consider the sum of high-frequency absolute returns, or the so-called "realized power" variation $\tilde{P}_{t+1,t}^{(m)}$, which is defined as $\sum_{j=1}^{m} |r_{t-(j-1)/m,t-(j-2)/m}|$. Regression (1.7) projects future realized volatility on past daily realized power.

Finally, to forecast the log of the quadratic variation, we consider log transformations of the five regressors in equations (1.3-1.7). In this fashion, our results would be directly comparable with those in the previous literature. To summarize, the MIDAS framework offers the ability to mix data sampled at different frequency, combined with a tightly parameterized model that allows different regressors to forecast volatility.

1.2   Comparison of MIDAS with Other Volatility Models

To further understand the flexibility of the MIDAS volatility models, it useful to compare them with other widely used models of conditional volatility, which for the purpose of

9

presentation will be written as:

$$\tilde{Q}_{t+H,t}^{(Hm)} = \mu_H^Q + \phi_H^Q \sum_{k=0}^{k^{max}} b_H^Q(k) \tilde{Q}_{t-k,t-k-1}^{(m)} + \varepsilon_{Ht}^Q \tag{1.8}$$

$$\tilde{Q}_{t+H,t}^{(Hm)} = \mu_H^G + \phi_H^G \sum_{k=0}^{k^{max}} b_H^G(k;\theta)(r_{t-k,t-k-1})^2 + \varepsilon_{Ht}^G \tag{1.9}$$

In equation (1.8), past $\tilde{Q}_{t,t-1}^{(m)}$ are used to predict $\tilde{Q}_{t+H,t}^{(Hm)}$ and the weights $b_H^Q(k)$ are not parameterized. When $H = 1$ such models for so called realized volatility, analyzed by Andersen et al. (2001, 2003), Andreou and Ghysels (2002), Barndorff-Nielsen and Shephard (2001, 2002a,b, 2003a), and Taylor and Xu (1997), often rely on Merton's (1980) arguments that arbitrarily accurate estimates of volatility can be obtained by increasingly finer sampling of returns. The above papers show that the use of high-frequency data is beneficial in predicting volatility. Again, when $H = 1$, the difference between (1.8) and (1.3) is the specification of the weights b_H^Q. In this regard, it is important to note that Andersen et al. (2003) advocate the use of long memory models to parsimoniously parameterize the weights. In particular, they consider models of the following type:

$$(1 - \sum_{k=1}^{5} b_A(k)L^k)(1 - L)^d \log \tilde{Q}_{t+1,t}^{(m)} = \mu + \varepsilon_t \tag{1.10}$$

Hence, using a fractional differencing approach one can capture with a single parameter d slowly decaying response patterns associated with long memory. In addition to the fractional differencing parameter d in equation (1.10) Andersen et al. (2003) advocate the use of an $AR(5)$ autoregressive expansion appearing on the left hand side of the equation. Hence, a total of 6 parameters (not including the constant) are used to model the autoregressive dynamics of realized volatility. Model (1.10) is expressed in terms of log volatility as Andersen et al. argue that the log transformation induces normality and therefore justifies the use of linear autoregressive models. This model will be our benchmark for all in-sample and out-of-sample forecast comparisons and is henceforth referred to as the "ABDL" model.

It is important to stress the differences between MIDAS regression models and the benchmark ABDL ARFI(5,d) model specification appearing in equation (1.10). None of the MIDAS regressions operate through autoregression whereas the ABDL specification follows much closer the tradition of ARCH-type models since $\tilde{Q}_{t+1,t}^{(m)}$ is projected onto its lagged values.

Furthermore, MIDAS regressions involve at most two parameters for the Beta polynomial, a scaling parameter and an intercept, i.e. less than the typical ABDL setting. The challenge is to outperform the ABDL specification while choosing: (1) the type of regressors; and (2) the decay patterns through judicious choice of parameterizations of the polynomial weighting schemes. The success of this challenge is the main argument of this paper.[7]

Finally, it should be noted that equation (1.9) is a version of the most widely used specification of conditional volatility, namely, the ARCH-type models of Engle (1982) (see also Bollerslev (1986)). In (1.9), future volatility is projected onto lagged daily squared returns (and a constant), where the weights are tightly parameterized via an autoregression such as in the popular GARCH(1,1) specification. Andersen et al. (2003) show that models appearing in equation (1.10) outperform ARCH-type models, which is why we use the former as a benchmark.

1.3   In-Sample Fit and Out-of-Sample Forecast Evaluations

The usefulness of the MIDAS framework is best judged by its ability to produce good in-sample fit and an out-of-sample forecasts of realized volatility. We elaborate here in some detail on how the various models will be compared. In the empirical results that follow, we consider the mean square error of a given MIDAS model, MSE^{MIDAS}, and compare it to the mean squared error from the ABDL model, denoted by MSE^{ABDL}. These measures are computed in-sample and in an out-of-sample experiment. Ultimately, we report the ratios MSE^{MIDAS}/MSE^{ABDL} as a measure of the forecasting performance of the MIDAS models relative to the ABDL. Ratios below 1 indicate that a given MIDAS model outperforms the benchmark. Since the ABDL has proven to be one of the most successful models in volatility forecasting, the MIDAS models have to clear a high mark.

We follow a slightly different approach for the in-sample and out-of-sample comparisons. Namely, the in-sample and out-of-sample comparisons different with respect to the variable being predicted.

In the out-of-sample comparison, we follow Andersen et al. (2003) who predict future log volatility, exponentiate the prediction, and then compute the square root to obtain

---

[7]In more recent work Andersen, Bollerslev and Diebold (2003) have advocated the use of so-called "HAR regressions," proposed in Corsi (2003) and sharing features with MIDAS regressions, involving regressors that exclude the jump component of quadratic variation. We will elaborate further on this later.

a prediction of future standard deviation. We use exactly the same strategy for our log volatility specification. For specifications where we predict the level of volatility (as opposed to the log), only a square root transformation is necessary. All out-of-sample MSE ratios reflect the relative performance of predict future standard deviations over four different horizons, ranging from one to four weeks.

The in-sample comparison is slightly different. The prediction of the MIDAS models is in terms of the left hand side specification, namely $V_{t+H,t}^{(Hm)}$. Hence, a MIDAS model predicting logs is compared with the ABDL model which predicts logs too. However, for the MIDAS regression predicting the level of volatility the ABDL model prediction is exponentiated. This may perhaps put the ABDL model at a slight disadvantage, yet that disadvantage is negligible.

It is important to emphasize another difference between ABDL models and MIDAS regressions. When predicting future, say, weekly volatility Andersen et al. (2003) predict the daily volatility for each day of the week and add up the forecasts. In a MIDAS regression the future weekly volatility is taken *directly* as target of interest in the prediction. This is a good illustration of the advantage of using a mixed-data sampling framework.

It is also worth noting that one might object to the use of MSE as a measure of forecast comparison. Arguably, this criterion may be criticized in the context of volatility predictions as for portfolio or risk management the objective function may neither be symmetric nor be directly related to a pure measure of statistical fit. Even as far as statistical criteria go, since we compare many models estimated with different number of parameters, one could argue in favor of the adjusted R^2 for comparisons. MIDAS regressions, however, always involve *less* parameters than the benchmark. Hence, penalizing a model for the number of parameters would actually penalize the benchmark model. We are therefore using the MSE criterion which effectively increases the hurdle for the performance of MIDAS regressions. Moreover, the in-sample MSE ratio of two models is equivalent to their unadjusted R^2 ratio. We opted for the ratio of the MSEs (rather than R^2s) because we can compare these measures across in-sample and out-of-sample results. Such a comparison can give us an idea of whether the in-sample results are being purely driven by over-fitting or sampling error. Finally, we do not test whether MSEs are statistically significant from each other, as this task is not so straightforward in this context.[8]

---

[8]See e.g. West (2004) for a discussion of the validity of asymptotic-based tests and the issues involved.

2 Results with Daily Data

We use the MIDAS framework to forecast the volatility in the US stock market. Our dataset consists of five-minute intra-day returns of the Dow Jones Composite Portfolio (DJ) over a ten year period, from April 1, 1993 to October 31, 2003. We also have five-minute returns for individual stocks that belong to the Dow. In this section, we focus exclusively on the DJ index and examine the evidence obtained from individual stocks later in the paper. Hence, the DJ index is the lead example, while evidence from individual stocks will be used to show the robustness of our findings. All return data are reported from 9:30 am to 4:05 pm every trading day. The returns for some days are removed from the sample to avoid the inclusion of regular and predictable market closures which affect the characterization of the volatility dynamics. In cleaning the dataset, we follow Andersen et al. (2001) who use a similar five-minute dataset with returns from the foreign exchange market. The final dataset contains 2,669 trading days with 80 observations per day for a total of 213,520 five-minute returns for each asset. From this dataset, we compute daily realized volatility, daily squared returns, daily absolute returns, daily ranges, and realized power.

As noted before, we consider four prediction horizons, from 1 through 4 weeks. The horizons cover most practically relevant cases, particularly in the context of Value-at-Risk computations, option pricing, and portfolio management. The regressions are run with non-overlapping $(\log)\tilde{Q}_{t+H,t}^{(Hm)}$, for $H = 5$ days $(1wk)$, $H = 10$ $(2wks)$, 15 $(3wks)$ and 20 $(4wks)$ days respectively. The MIDAS regressions are obtained with the Beta lag structure (1.2). We use a truncation of 50 daily lags in the estimation, or $k^{max} = 50$ since using longer lags has little effect on the results. The models are estimated with quasi-maximum likelihood and the standard errors are obtained using the Bollerslev and Wooldridge (1992) approach to account for heteroscedasticity. A detailed discussion about estimating MIDAS regression models can be found in Ghysels et al. (2002b). In general, the errors of MIDAS regressions may feature autocorrelation or heteroskedasticity. However, in the current application, all regressions involve non-overlapping prediction samples to avoid autocorrelation in the residuals due to overlapping prediction horizons. Regression diagnostics, discussed later, will reveal that the residuals indeed appear uncorrelated and in some cases Gaussian as well.[9]

---

[9]For the ABDL benchmark model we re-estimate the AR parameters, whereas the fractional differencing parameter is pre-set at 0.40096. The latter parameter is quite stable across assets, as noted by Andersen et al. (2003). We therefore follow their strategy of keeping a common value of d across all assets.

2.1 In-Sample Fit

Table 1 presents the in-sample ratios MSE^{MIDAS}/MSE^{ABDL}. The results from forecasting the two measures of volatility $\tilde{Q}_{t+H,t}^{(Hm)}$ and $\log\left(\tilde{Q}_{t+H,t}^{(Hm)}\right)$ are presented in Panels A and B, respectively. The results from the April 1993 to March 2001 subsample and for the entire sample are presented. In Panel A of Table 1, we look at the forecasts of $\tilde{Q}_{t+H,t}^{(Hm)}$ using various daily predictors over the April 1993 to March 2001 subsample and the full sample ending in 2003. When we use realized daily variances as regressors (equation (1.3)), the relative MSE ranges from 0.871 to 0.683 across the forecasting horizons and samples. The results are very much similar across the full sample and the subsample. The log transformation puts the MIDAS and ABDL models more on an equal footing (in-sample), as noted from Panel B in Table 1. The MSE ratio varies from 0.924 to 0.812, again depending on horizon and length of sample, with a more pronounced advantage for the MIDAS models at the longer horizons.

One difference between the MIDAS regressions and the benchmark is the weighting scheme. When we use lagged realized volatility as a MIDAS predictor, there is no other difference between the two models. In the ABDL model, the weight on past daily observations is captured by the ARFI polynomial, whereas in the MIDAS framework, it is given by the Beta polynomial. Figures 2 and 3 show the profiles of the estimated MIDAS regression weights (entire sample) for the $4wks$ horizon for the $\log\left(\tilde{Q}_{t+H,t}^{(Hm)}\right)$ and $\tilde{Q}_{t+H,t}^{(Hm)}$ regressions, respectively. We can directly compare the weights of the ABDL specification and the corresponding log volatility MIDAS model. Figure 2 displays these two models and several other MIDAS regressions which will be discussed in detail later. The weights in the figures are normalized to sum up to one and hence have a common scale. Consequently, the figures tell us the relative share of weights across all the lags. For the moment, we focus on the patterns for the ABDL model (solid line) and the QV MIDAS model as it is labeled in the figure. The differences are remarkable. The ABDL weighting scheme is somewhat jagged for the first lags due to the AR(5) augmentation and then we recognize the long slowly decaying tail typical for long memory models. In fact, the weights persist beyond 50 days. In contrast, the QV MIDAS model produces a smoothly declining weighting scheme that dies out roughly around 25 to 30 days. Hence in the MIDAS regressions more weight is attributed to the more recent past. This explains why there are no gains made with increasing the lag length beyond 50 days.[10]

---

[10]In a related paper, pertaining to the risk-return trade-off (Ghysels et al. (2002a)) it is shown (see for instance Figure 3) that a monthly GARCH has exactly the same problem, namely the weights die out too

Figure 3 displays similar patterns (without the ABDL specification) for MIDAS regressions not involving the log transformation. The memory features are, not surprisingly, similar. Returning to the log transformation, it is also worth noting that log squared returns (which are identical to log absolute returns up to a scaling factor) have a decay pattern that resembles more closely that of the ABDL specification.[11] Obviously, squared returns perform poorly relative to the ABDL specification, as daily squared returns are a noisy measure of past realized volatility. In the language of MIDAS, squared returns are poor regressors with a decay pattern (obtained via Beta polynomial weight fitting) similar to the ARFI(5,d) ABDL specification.

Table 2 complements Figures 2 and 3. The table reports regression diagnostics and numerical values for some of the MIDAS regressions. The parameter estimates of θ are not displayed because they have no economic interpretation in and of themselves. We show the other parameter estimates including a summary of the estimated weight function. In addition, we include the Q Portmanteau test for serial correlation in the residuals. The Q test shows that, except for the short horizon, all residuals appear uncorrelated. This is of course an appealing feature for the MIDAS regressions. Although not reported, we also tested the residuals for normality and found, using the Jarque-Bera test, that the residuals of the log volatility MIDAS regression appear normal for all regressors. For the level volatility MIDAS regression models, again using the Jarque-Bera test, the null of normality was barely rejected. If we look at the weight function results, the $day1$ weight is much higher with the log specification compared to the level. The log specification also shows higher weights throughout as it is more predictable. From the table, we can also see that between 80 and 99 percent of the weight is placed on the first twenty daily lags of the regressors. However, between 40 and 50 percent is placed on lagged daily forecasters between days 6 and 20. Hence, a daily model that takes into account only, say, five daily lags is likely to miss a lot of the dependence in the data. As another remark to Table 2, the MIDAS regressions not involving past returns may result in negative intercepts. (Naturally, the problem is not present for the log specification) There is obviously a scaling issues that get resolved through the regression but it does not

---

slowly when compared to a monthly MIDAS (the daily GARCH in the same figure yields the opposite result but this is a more complex comparison since it involves aggregation - it is therefore not directly related to what is reported here).

[11]It also is interesting to note that in the context of the risk-return tradeoff Ghysels et al. (2002a) obtain estimates of volatility based on squared daily returns, since high-frequency data are not available for long historical samples. In the context of the risk-return tradeoff with monthly returns they also find long lags using daily squared returns, as depicted in Figure 2.

prove to adversely affect the out-of-sample performance of the model, as shown below.

So far, we discussed MIDAS models that are almost directly comparable to the ABDL benchmark. Perhaps the most interesting MIDAS results involve the performance of daily range and realized power (equations (1.6) and (1.7)). Table 1 presents the in-sample ratios MSE^{MIDAS}/MSE^{ABDL} for both models. Forecasts with $[hi-lo]_{t-k,t-k-1}$ produce relative MSEs in the range of 0.722 and 0.993. It appears also that the forecasting performance of the MIDAS regressions in the subsample is slightly better than in the full sample. Overall the ratios are lower than the goodness of fit obtained from $\tilde{Q}^{(m)}_{t-k,t-k-1}$, $(r_{t-k,t-k-1})^2$, and $|r_{t-k,t-k-1}|$ at any horizon. The superior performance of ranges has been noted before, see e.g. Gallant et al. (1999).[12] One advantage of using $[hi-lo]_{t-k,t-k-1}$ is that it arguably does not involve any measurement error issues (see e.g. Alizadeh et al. (2001)), Andersen et al. (2003), Andreou and Ghysels (2002), Barndorff-Nielsen and Shephard (2001, 2002a,b, 2003a) for further discussion of measurement error issues).

The most remarkable predictor of volatility is the realized power $\tilde{P}^{(m)}_{t-k,t-k-1}$. At various horizons, the MSE ratios obtained from the realized power MIDAS regressions are between 0.606 and 0.826 for the level specification and 0.761 and 0.921 for the logs. Particularly at longer horizon, the in-sample performance is considerably better than those obtained with $[hi-lo]_{t-k,t-k-1}$, and significantly better than the other predictors. Hence, the best predictor of future increments in quadratic variation is the past realized power measure.

Turning to Table 2, the weights attributed to the different lags are slightly larger when compared to those from the quadratic variation specification. In Figures 2 and 3, we observe that the decay patterns across the various MIDAS regression specifications (excluding daily squared and absolute returns) are quite similar. This evidence suggests that there are two reasons for the success of the MIDAS approach, namely: (1) the weighting scheme; and (2) the judicious choice of regressors. Both factors are important contributors in the success of the MIDAS forecasts.

As a final comment, it is interesting to compare the weight functions across forecasting

---

[12]The Gallant et al. (1999) approach to range-based volatility prediction shares some features with the regression appearing in (1.6). The main difference is that the setup in Gallant et al. (1999) involves a parametric continuous time model and a so called re-projection procedure using simulated samples where parameter proliferation is less of an issue. The use of MIDAS regressions allows us to treat parsimoniously the predictive power of range-based measures in a reduced-form setting without specifying an underlying structural continuous time data generating process for returns.

horizons.[13] At a short horizon of 1 week, a larger fraction of the MIDAS weights are placed on the first daily lag (between 13 and 20 percent, depending on the specification) and very little weight is placed on lags beyond the first 20 days (between 0.3 and 3 percent). As the horizon increases to 4 weeks, the weight on the first day decreases substantially and a significant weight (from 10 to 20 percent) is placed on lags beyond the first 20 days.

2.2 Out-of-Sample Comparisons

The in-sample fit measures indicate that the MIDAS models are flexible enough to capture the features of the data. However, the ultimate test of forecasting models is whether they maintain their performance out-of-sample. The out-of-sample MSEs of the MIDAS models relative to the benchmark are displayed in Table 3 which features a few interesting results.

First, the realized power $\tilde{P}_{t-k,t-k-1}^{(m)}$ is the dominant out-of-sample predictor of future realized volatility. Also, MIDAS regressions involving range and past daily realized volatility continue to outperform the ABDL benchmark model. These findings parallel the in-sample results.

Second, for the realized power specification, the gain in MSE is typically of the order of 20 to 30 percent. Hence, the in-sample MSEs (Table 1) and out-of-sample MSEs (Table 3) are very comparable in magnitude. This indicates that the performance of the MIDAS model is not due to in-sample over-fitting.

Third, longer horizon forecasts do not necessarily produce worse out-of-sample performance (at least up to 4 weeks). This is because we use the mixed-frequency regression to directly produce long horizon forecasts instead of cumulating short horizon forecasts. Moreover, we saw that the estimated MIDAS weights are different as the horizon increases, which suggests that our framework is flexible enough to accommodate various volatility dynamics.

Finally, the MIDAS regressions predicting the log and the level of volatility tend to produce similar out-of-sample forecasts. Hence, despite the extra transformation (exponentiation) and its potential source of bias, it seems that the log transformation is a viable way of formulating forecasts.[14]

---

[13]We thank the Referee for suggesting this analysis.

[14]We have also experimented with MIDAS models predicting directly the square root of $\tilde{Q}_{t+H,t}^{(Hm)}$. They did not perform as well as the log specification, indicating the important role played by the predictability rather than bias.

2.3 Robustness of Findings: Evidence from Individual Stocks

In this section, we examine the robustness of our findings. To do so, we consider evidence from a collection of individual stocks that belong to the DJ Index. The stocks considered are: Disney (DIS), Exxon Mobil (XOM), General Electric (GE), Honeywell (HON), J.P. Morgan Case and Co (JPM) and McDonald's (MCD). These six stocks were part of the DJ Index over the entire sample and represent different sectors of the US economy. The results appear in Table 4. To avoid reporting many results, we simply focus on the out-of-sample performance.

Table 4 is analogous to Table 3 for the DJ index. The results obtained with the DJ index remain true for the individual stocks. The realized power variation produces the best out-of-sample forecasts among all level $\tilde{Q}_{t+H,t}^{(Hm)}$ MIDAS regressions. This is true for all stocks and across horizons. For the log specification, the realized power also dominates all other predictors. For a few stocks and for selected horizons, the range and the quadratic variations produce marginally better forecasts. The overall results from the volatility of six major stock returns convey the same message. First, the MIDAS weights are flexible enough to produce in- and out-of-sample volatility forecasts that match those from the most benchmark. Second, the realized power volatility is the best predictor of future realized volatility.

As a final robustness note, in a previous version of the paper we obtained very similar results using a different high-frequency dataset from the foreign exchange market.[15] Unfortunately, that dataset was outdated and our efforts to update it and to provide new out-of-sample results were unsuccessful. However, given that the predictive accuracy of the MIDAS regressions with realized daily power predictors has been demonstrated in- and out-of-sample, with an index and individual data returns, and across markets leads us to believe that our results are indeed quite robust.

3   MIDAS Regressions and High-Frequency Data

So far, we discussed that MIDAS regressions are able to accommodate a variety of predictors in a parsimonious representation of long lag structures. In this section, we explore another appealing feature of the MIDAS framework – its ability to easily mix sampling frequencies.

---

[15]We were using the Olsen and Associates 5-minute DM/US $ return data covering the period December 1986 to November 1996.

More concretely, we investigate whether the direct use of high-frequency data results in better forecasts of future volatility. We revisit the choice of regressors in the context of high-frequency (5-minute) data and discuss information sets, polynomials, and implicit filters. The final subsection explores the empirical evidence of volatility predictability using high-frequency data.

3.1 Choice of Intra-Daily Regressors

Let us reconsider the argument that arbitrarily accurate estimates of volatility can be obtained by increasing the sampling frequency of returns. Unlike equation (1.3), where future increments in quadratic variation are projected onto a number of lags of daily quadratic variation, we explore a new dimension inherent to MIDAS regression. Instead of aggregating intra-daily squared returns or absolute returns to daily measures $\tilde{Q}_{t,t-1}^{(m)}$ and $\tilde{P}_{t,t-1}^{(m)}$ why not run MIDAS regressions directly on the high-frequency data? Using the notation $B(L^{1/m})$ for a distributed lag polynomial in $L^{1/m}$ which has the property $L^{1/m}r_{t,t-1/m} = r_{t-1/m,t-2/m}$, we consider the regressions:

$$
\begin{align}
\tilde{Q}_{t+H,t}^{(Hm_1)} &= \mu_H^s + \phi_H^s B_H^s(L^{1/m_2})[r_{t,t-1/m_2}]^2 + \varepsilon_{Ht}^s \tag{3.1} \\
\tilde{Q}_{t+H,t}^{(Hm_1)} &= \mu_H^a + \phi_H^a B_H^a(L^{1/m_2})|r_{t,t-1/m_2}| + \varepsilon_{Ht}^a \tag{3.2}
\end{align}
$$

where the first equation involves projecting $\tilde{Q}_{t+H,t}^{(Hm_1)}$ onto past squared returns sampled at frequency $1/m_2$. Note that the sampling frequency used to compute quadratic variation (m_1) may differ from the sampling frequency used to compute high-frequency returns. The specification appearing in (3.1) is closely related to both the daily realized volatility model in (1.8) and the daily MIDAS regression in (1.3). With, say, 50 days of lags there is a direct comparison between (3.1) and (1.3) since they both involve the same information set. Similar arguments apply to absolute intra-daily returns used in equation (3.2) and the MIDAS regression involving realized powers in equation (1.7). We now turn to topics pertaining to information sets and polynomials in the next two subsections.

3.2 Information Sets

We will consider first projections involving squared returns, namely:

$$E_L(\tilde{Q}_{t+H,t}^{(Hm_1)}|[r_{t,t-j/m_2}]^2, j \geq 0) \equiv \mu_H^s + \phi_H^s B_H^s(L^{1/m_2})[r_{t,t-1/m_2}]^2 \qquad (3.3)$$

where $E_L(.|.)$ denotes a linear projection. When we consider information sets involving past realizations of measured quadratic variation, i.e. $\tilde{Q}_{t-j,t-j-1}^{(m)}$, $j = 1, \ldots$ we obtain:

$$E_L(\tilde{Q}_{t+H,t}^{(Hm_1)}|\tilde{Q}_{t-j,t-j-1}^{(m_1)}, j \geq 0) \equiv \mu_H^Q + \phi_H^Q B_H(L)\tilde{Q}_{t,t-1}^{(m_1)}. \qquad (3.4)$$

Since the information set $[\tilde{Q}_{t-j,t-j-1}^{(m)}, j \geq 0]$ is a subset of $[r_{t,t-j/m}]^2, j \geq 0$ (assuming a common m) we know the following property holds for linear projections and quadratic loss functions:

$$MSE(E_L(\tilde{Q}_{t+H,t}^{(m)}|[r_{t,t-j/m}]^2, j \geq 0)) \leq MSE(E_L(\tilde{Q}_{t+H,t}^{(m)}|\tilde{Q}_{t-j,t-j-1}^{(m)}, j \geq 0)) \qquad (3.5)$$

The reduction of information from $[r_{t,t-j/m}]^2, j \geq 0$ to $[\tilde{Q}_{t-j,t-j-1}^{(m)}, j \geq 0]$ has the resulting (linear) projection only in terms of daily lags, whereas the linear projection (3.1) involves many lags of high-frequency data. On the one hand, the relation in (3.5) may not be attained in practice as the linear projection in (3.3) involves a very large number of parameters. On the other hand, a parsimonious model built on a parametric specification of the high-frequency dynamics may yield very complicated projection formulas. Mixed data sampling regressions exploit a much larger information set, without the cost of parameter proliferation, which is a defining feature of mixed data sampling regression models. Hence, we try the keep the rich information set without being penalized by prohibitively large parameter spaces.

It should also be noted that the MIDAS regression approach is explicit about the information sets involved. Namely, if one considers two sampling frequencies m_1 and m_2 with, for instance, $m_1 < m_2$, then there is a MIDAS regression model associated with each one of the sampling frequencies. Such MIDAS regressions, involving linear projections $E_L(\tilde{Q}_{t+H,t}^{(Hm_1)}|[r_{t,t-j/m_1}]^2, j \geq 0)$ and $E_L(\tilde{Q}_{t+H,t}^{(Hm_1)}|[r_{t,t-j/m_2}]^2, j \geq 0)$, could be viewed as nested models, though this may not lead to any practical testable restrictions. However, we can use standard regression methods to appraise the fit of the two projections. Namely, in practice we may run into microstructure noise and linear regressions with m too high may therefore

under-perform. The MIDAS regression framework and the explicit treatment of information sets allows us to address these issues rather directly, instead of assessing separately first the quality of realized volatility extractions and then proceed in a second stage with model specification.

The arguments regarding information sets also apply to absolute returns and realized power specifications. Consider the following linear projection:

$$E_L(\tilde{Q}_{t+H,t}^{(Hm_1)}||r_{t,t-j/m_2}|, j \geq 0) \equiv \mu_H^a + \phi_H^a B_H^a(L^{1/m_2})|r_{t,t-1/m_2}| \tag{3.6}$$

and the linear projection onto past realized absolute power defined as $\tilde{P}_{t-j,t-j-1}^{(m)}$, yielding:

$$E_L(\tilde{Q}_{t+H,t}^{(Hm_1)}|\tilde{P}_{t-j,t-j-1}^{(m_1)}, j \geq 0) \equiv \mu_H^P + \phi_H^P B_H^P(L)\tilde{P}_{t,t-1}^{(m_1)} \tag{3.7}$$

The arguments regarding information sets apply to these two projections as well.[16]

3.3 Polynomials and Implicit Filters

The daily regression (1.3) involving quadratic variation can be viewed as factorizing the polynomial $B(L^{1/m})$ in equation (3.1) into subpolynomials that are directly interpretable. For example, in the linear projection (3.3) one could consider a factorization of the following type: $B(L^{1/m}) \equiv B_d(L)B_{rv}(L^{1/m})$, where $B_d(L)$ is a polynomial in lag L (hence involving daily sampling) and $B_{rv}(L^{1/m})$ is a predetermined polynomial representing realized volatility, i.e. $B_{rv}(L^{1/m}) \equiv \sum_{j=1}^{m}(L^{1/m})^j$ which results in taking the sum of $1/m$ sampled squared returns and hence computing $Q_{t,t-1}^{(m)}$. As noted before, and discussed at length notably by Andreou and Ghysels (2002b) and Barndorff-Nielsen and Shephard (2001, 2002a,b, 2003a), realized volatilities are estimates of the quadratic variation of a diffusion subject to measurement error. The presence of measurement error has prompted Andreou and Ghysels (2002b) to consider rolling sample estimators of quadratic variation, called HQV, referring to historical quadratic variation. MIDAS regression volatility models also include factorizations involving HQV estimators. For example, one could consider a factorization of the following type: $B(L^{1/m}) \equiv B_d(L)[B_w(L)B_{rv}(L^{1/m})]$, where $B_w(L)$ represents a rolling

---

[16]It should be noted however, that there are differences in the asymptotics, namely when $m \to \infty$. With $\tilde{P}$ we need a $\sqrt{m}$ normalization, whereas with $\tilde{Q}$ there is no need to normalize by m as shown by Barndorff-Nielsen and Shephard (2003b, 2004). We do not endeavor into the asymptotics of $m \to \infty$ in the context of this paper.

sample predetermined polynomial combining daily realized volatilities. Alternatively, one could also consider the intra-daily smoothing proposed by Andreou and Ghysels (2002b), which consists of taking $B_d(L)[B_w(L^{1/m})B_{rv}(L^{1/m})]$, with weighting schemes rolling through intra-daily quadratic variation estimators. It may be clear by now that the direct estimation of MIDAS regression volatility models, is a simple approach to accomplish two operations in one single step. The first operation is filtering past high-frequency returns such that they represent past daily volatilities. The second operation is the projection of future realizations onto past filtered volatilities to obtain predictions. The combination of the two operations into one step has many advantages. For instance, with a flexible enough lag structure, one can compute predictions that exploit directly the high-frequency data set. Combining high-frequency data first into daily estimates entails a loss of information that cannot be recovered in the second step. Moreover, the design of extraction filters, i.e., is the choice of quadratic variation estimator, is by-passed and therefore the worry about noisy volatility estimates is avoided.

3.4   Empirical Results with High-Frequency Data

We now turn our attention to the estimation of MIDAS regressions involving intra-daily (5-minute) data. A major issue is that intra-daily seasonal patterns in financial markets have been widely documented in the literature since the work of Wood et al. (1985). Clearly tomorrow's quadratic variation does not feature seasonal intra-daily fluctuations, whereas the high-frequency intra-daily returns do. The most practical way to address such a situation is to remove the high-frequency intra-daily seasonal patterns from returns prior to fitting the lag structures. One may therefore wonder whether one could improve MIDAS regressions involving intra-daily data by considering seasonal adjustment prior to fitting a MIDAS polynomial. On the other hand, one could also accommodate the intra-daily patterns directly via the specification of the polynomial lag structure. A mixture of Beta lag polynomials can accommodate this by centering one polynomial on the mean of a seasonal lag. For example, suppose one day's worth of lags is 288, then there are 24 hourly peaks in the lag structure. Therefore one can take $\theta = ((a_i, b_i), i = 1, \ldots, 24)$, and center the Beta polynomials at the hourly lags, i.e. $a_i/(a_i + b_i) = 12i$ for $i = 1, \ldots, 24$. This would not be very appealing, however, as one would involve at least 48 parameters, to obtain a polynomial lag structure flexible enough to accommodate all the ups and downs of periodic market activity throughout a business day. Note that in less complex seasonal settings, involving quarterly or monthly

data, the strategy may indeed be appealing. The alternative is to seasonally adjust the data prior to fitting MIDAS regressions, a subject to which we turn next.

3.4.1 Seasonality in MIDAS Regressions

There are theoretical arguments that support seasonally adjusting the data prior to fitting a MIDAS regression model, and those arguments are spelled out in detail in Hansen and Sargent (1993) and Sims (1993). We consider again a generic regressor $x^{(m)}$, which may be squared returns, absolute returns or range, that is adjusted for (intra-daily) seasonal patterns. We denote by $x^{(m)}$ the raw regressors, and $\tilde{x}^{(m)}$ the adjusted ones. We have considered two alternative schemes for removing the periodic patterns. The first consists of subtracting hourly means of squared returns, i.e. $[\tilde{x}_{t-j/m}^{(m)}]^2 \equiv [\tilde{x}_{t-j/m}^{(m)}]^2$ - $M_{h(j)}$, where $h(j)$ is the hour to which the jth observation belongs and M_h is the mean of $[x_{t-j/m}^{(m)}]^2$ for hour h. The second scheme is slightly more general as it adds a day-of-the-week effect, namely $[\tilde{x}_{t-j/m}^{(m)}]^2 \equiv [x_{t-j/m}^{(m)}]^2$ - $M_{h(j),d(t)}$, where $d(t)$ is the day of the week of day t and $M_{h,d}$ is the mean of $x_{t-j/m}^{(m)}]^2$ for hour h on day d. Obviously one could consider other adjustment schemes as well.[17]

Let us consider the example of squared return regressors and rewrite (3.1) in terms of adjusted returns as

$$\tilde{Q}_{t+H,t}^{(Hm_1)} = \mu_H^s + \phi_H^s B_H^s(L^{1/m_2})([\tilde{r}_{t,t-j/m_2}]^2 - M_{h(j),d(t)}) + \varepsilon_{tH}^s$$

which can also be rewritten as

$$\tilde{Q}_{t+H,t}^{(Hm_1)} = \mu_H^s + \phi_H^s B_H^s(L^{1/m_2})[\tilde{r}_{t,t-j/m_2}]^2 + \beta_2 z_t + \varepsilon_{tH}^s \tag{3.8}$$

where $\beta_2 z_t$ represents a set of seasonal dummies, i.e. $z_t \equiv (z_t^d, d = 1, \dots, 5)$ with $z_t^d = 1$ if $d = d(t)$ and zero otherwise. Finally, the parameters β_2 are coefficients implicitly defined by the relationship $B_1(L^{1/m})M_{h(j),d(t)} \equiv \beta_2 z_t$. Comparing equations (3.8) and (3.1) yields the result that in principle one could run a regression with seasonal dummies and raw data to estimate $B_H^s(L^{1/m})$, instead of adjusting squared returns first and estimate the polynomial lag structure with $r_t^{(m)}$. Unfortunately, both are not necessarily equivalent even in large

---

[17]There is lots of evidence that the intra-daily seasonality is best captured by a multiplicative type of factor, see e.g. Martens et al. (2002). This suggests that division by a seasonal factor, rather than demeaning, is a natural approach to seasonal adjustment. In the log volatility specifications involving log regressors we de facto apply such a multiplicative type adjustment.

samples. The main reason is that we are dealing with potentially misspecified models since the polynomial lag structure is an approximation to a highly over-parameterized projection. In such circumstances, it is often better to adjust series prior to fitting "approximate" models. The formal arguments in Hansen and Sargent (1993) and Sims (1993) are based on the fact that the approximation error may be unduly influenced by the large seasonal variation in the data. In the empirical work involving intra-daily data, we show that adjusting the squared returns prior to running mixed data sampling regressions does not seem to have much of an impact. On the contrary, using raw data yields better results that those obtained with adjusted returns.

3.4.2 Empirical Results

Table 5 presents the in-sample ratios MSE^{MIDAS}/MSE^{ABDL}, where the MIDAS regressions involve either unadjusted returns or hourly adjusted returns, as discussed in the previous subsection. We report again results from forecasting the two measures of volatility $\tilde{Q}_{t+H,t}^{(Hm)}$ and $\log\left(\tilde{Q}_{t+H,t}^{(Hm)}\right)$ for the April 1993 to March 2001 subsample and for the entire sample. Note that the log specification is reduced to one model as the regressors are identical apart from a scaling factor. With unadjusted intra-daily squared returns we have in-sample fits that are roughly *equal* to those obtained with the aggregated measure $\tilde{Q}_{t,t-1}^{(m)}$. This is true for both the subsample and the full sample. Likewise, with unadjusted intra-daily absolute returns we have in-sample fits that are also roughly equal to those obtained with the corresponding aggregated measure $\tilde{P}_{t,t-1}^{(m)}$. The log specification, for which there is only one regression, also comes close to daily regression MIDAS models with $\tilde{Q}_{t,t-1}^{(m)}$, rather than $\tilde{P}_{t,t-1}^{(m)}$. Perhaps this is not so surprising as the log transformation puts the squared and absolute returns on an equal footing.

An interesting observation emerges from looking at the right panel of Table 5. Hourly adjusting the returns does not yield any gains in the in-sample fit. In fact, quite often it yields slight deteriorations of the in-sample fit compared to the MIDAS regressions with unadjusted returns (and therefore also with the corresponding daily regressors). Note that in none of the cases any explicit penalty was introduced for the use of parameters in estimating the adjustments. This finding suggests that more complicated models may be needed to adjust for seasonality in returns, see e.g. Andersen and Bollerslev (1997) for further discussion on this matter. However, more complicated adjustment models may also defy the purpose of

24

parsimony. In a different context, the finding may not be true, as there are probably many other settings where seasonals are easier to model. Macroeconomic data series would be a good example to investigate this question further, as a monthly cycle would be easier to capture than an intra-daily one. The findings reported in Table 5 are significant, however, as one advantage of computing daily measures is that seasonality issues are foregone through the aggregation process. It appears from our findings that the presence of seasonals is not an obstacle to the application of MIDAS models to high-frequency financial data.

Figure 4 displays the profile of polynomials estimated with seasonally adjusted and unadjusted returns. This figure plots the weights that the MIDAS estimators of one week ($1wk$) prediction horizon for a MIDAS model predicting log increments of quadratic variation based on log 5-minute high-frequency intra-daily squared (absolute) returns appearing in regression (3.1). The MIDAS polynomial weights for both hourly adjusted and unadjusted squared returns are displayed. We note that there is very little difference between the polynomials, except for the fact that the MIDAS regression involving unadjusted returns tends to put more weight on the most recent days, in comparison to the MIDAS regression involving adjusted returns.

The out-of-sample results appear in Table 6. The results parallel the in-sample findings, namely the direct use of unadjusted intra-daily high-frequency data neither helps nor hurts. Hourly adjustment do again much worse and are therefore not recommendable.

Overall, we note that MIDAS regressions can handle directly intra-daily high-frequency data. The advantage of our unified framework is that it enables us to conclude that the incremental value of using directly high-frequency data appears to be non-existent, at least in the present context of 5-minute equity returns.

4   Conclusions

We study the predictability of return volatility with MIDAS regressions. Our approach allows us to compare forecasting models with different measures of volatility, frequencies, and lag lengths. While the main focus of this paper is volatility forecasting, it is clear that the MIDAS framework is general in nature and can find a good use in any empirical investigation that involves data sampled at different frequencies. Simplicity, robustness, and parsimony are three of its main attributes.

We report several intriguing findings regarding the predictability of weekly to monthly realized volatility in equity markets. First, we find that daily realized power outperforms daily realized volatility and that daily and intra-daily absolute returns outperform respectively daily and intra-daily squared returns. This set of results suggests that absolute returns are very successful at capturing fluctuations in future return volatility, despite the predominant emphasis in the literature on squared returns. Also, we find that daily ranges are extremely good forecasters of future volatility and are only second to realized power. This last finding is consistent with results in Gallant et al. (1999), Alizadeh et al. (2002) and Engle and Gallo (2003), among others, who use different methods and different data. Finally, we show that the direct use of high-frequency data does not necessarily lead to better volatility forecasts.

Our paper leaves one unanswered issue: Why is realized power such a good predictor of future volatility? While there still isn't a satisfactory answer to this question, several recent papers have made considerable progress. First, there is now an elegant asymptotic theory that was developed in a set of papers by Barndorff-Nielsen and Shephard (2003b, 2004) showing how realized power, and a measure called bi-power, exclude the jump component of increments in quadratic variation. Building on the Barndorff-Nielsen and Shephard analysis, Forsberg and Ghysels (2004) study the theoretical properties of realized power in the context of continuous time stochastic volatility models. They confirm the intuition that realized power is a persistent process more closely related to the long run component of volatility, which explains its success as a regressor in the context of MIDAS regressions. Forsberg and Ghysels (2004) also show that realized power has far less measurement noise in comparison with realized volatility and bi-power. They also discuss other appealing features of realized power, such as higher persistence and predictabiity in comparison with bi-power and realized volatility.

In a recent paper, Andersen, Bollerslev and Diebold (2003) suggest to use the bi-power variation as predictor of volatility because jump components are "noise" not helpful in predicting future volatility . The bi-power regressor is used in "HAR regressions," proposed in Corsi (2003), and is shown to improve the forecasting of volatility. Our paper shares a common theme with Andersen, Bollerslev and Diebold (2003). Both papers argue in different ways that taking a volatility measure immune to jumps - such as realized power or bi-power, is a good regressor to predict future volatility.[18]

---

[18]Early drafts of the current paper and versions of the Andersen, Bollerslev and Diebold (2003) paper

We have kept the mixed data sampling regressions as simple as possible in the interest of clarity and conciseness. However, there are a host of issues, such as asymmetries, multiple predictors, and option-implied measures of volatility, that merit further attention. These extensions are easily accommodated in the MIDAS framework and some of them are addressed in Ghysels et al. (2002a). Others remain to be explored.

were written almost simultaneously and independently.

References

[1] Aït-Sahalia, Y., Mykland, P. A. and Zhang, L. (2003), "How often to sample a continuous-time process in the presence of market microstructure noise", Working Paper w9611, NBER.

[2] Alizadeh, S., M. Brandt and F. X. Diebold, (2002), "Range-based estimation of stochastic volatility models", *Journal of Finance*, 57, 1047-1091.

[3] Andersen, T. and T. Bollerslev (1997), "Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long Run in high-frequency Returns", *Journal of Finance*, 52, 975-1005.

[4] Andersen, T. and T. Bollerslev (1998), "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts", *International Economic Review*, 39, 885-905.

[5] Andersen, T., T. Bollerslev, F. X. Diebold and P. Labys (2001), "The Distribution of Exchange Rate Volatility", *Journal of American Statistical Association*, 96, 42-55.

[6] Andersen, T.G., Bollerslev, T. and F.X. Diebold, (2002), " Parametric and Nonparametric Volatility Measurement," in L.P. Hansen and Y. Aït-Sahalia (eds.), Handbook of Financial Econometrics . Amsterdam: North-Holland, forthcoming.

[7] Andersen, T., T. Bollerslev, F. X. Diebold and P. Labys (2003), "Modeling and Forecasting Realized Volatility", *Econometrica*, 71, 529-626.

[8] Andersen, T.G., Bollerslev, T. and F.X. Diebold, (2003), "Some Like it Smooth, and Some Like it Rough: Untangling Continuous and Jump Components in Measuring, Modeling, and Forecasting Asset Return Volatility," Manuscript, Duke, Northwestern and University of Pennsylvania.

[9] Andreou, E. and E. Ghysels (2002), "Rolling sample volatility estimators: some new theoretical, simulation and empirical results" *Journal of Business and Economic Statistics*, 20, 3, 363-376.

[10] Baillie, R.T., T. Bollerslev and H. O. Mikkelsen (1996), "Fractionally integrated generalized autoregressive conditional heteroskedasticity" *Journal of Econometrics*, 74, 3-30.

[11] Bandi, F.M. and J.R. Russell,(2003), "Microstructure noise, realized volatility, and optimal sampling", Working paper, Graduate School of Business, The University of Chicago.

[12] Barndorff-Nielsen, O. and N. Shephard (2001), "Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion), *Journal of the Royal Statistical Society*, Series B, 63, 167-241.

[13] Barndorff-Nielsen, O. and N. Shephard (2002a), "Econometric analysis of realized volatility and its use in estimating stochastic volatility models", *Journal of the Royal Statistical Society*, Series B, 64, 253-280.

[14] Barndorff-Nielsen, O. and N. Shephard (2002b) "Estimating quadratic variation using realised variance" *Journal of Applied Econometrics*, 17, 457-477.

[15] Barndorff-Nielsen, O. and N. Shephard (2003a) "How accurate is the asymptotic approximation to the distribution of realised volatility?" in D.W.K. Andrews, J. Powell, P. Ruud and J. Stock (ed.), *Identification and Inference for Econometric Models. A Festschrift for Tom Rothenberg*, Cambridge University Press, (forthcoming).

[16] Barndorff-Nielsen, O. and N. Shephard (2003b) "Realised power variation and stochastic volatility" *Bernoulli* 9, 243-265.

[17] Barndorff-Nielsen, O. and N. Shephard (2004) "Power and bipower variation with stochastic volatility and jumps" (with discussion) *Journal of Financial Econometrics*, 2, 1-48.

[18] Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, 31, 307-327.

[19] Bollerslev, T., R. Engle and D. Nelson (1994), "ARCH Models, in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Vol. 4, Elsevier Science.

[20] Bollerslev, T. and J. M. Wooldridge (1992), "Quasi-Maxmimum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances," *Econometric Reviews*, 11, 143172.

[21] Bollerslev, T. and J. Wright (2001), "High-frequency data, frequency domain inference, and volatility forecasting", *Review of Economics and Statistics*, 83, 596-602.

[22] Corsi, F. (2003), "A Simple Long Memory Model of Realized Volatility", Manuscript, University of Southern Switzerland.

[23] Davidian M. and R.J. Carroll (1987), "Variance Function Estimation", *Journal of the American Statistical Association*, 82, 1079-1091.

[24] Dhrymes, P. (1971), *Distributed Lags: Problems of Formulation and Estimation*, Holden-Day, San Francisco.

[25] Ding, Z., and C. W. J. Granger (1996), "Modeling Volatility Persistence of Speculative Returns: A New Approach", *Journal of Econometrics*, 73, 185-215.

[26] Ding, Z., C. W.J Granger and R. F. Engle (1993), "A long memory property of stock market returns and a new models", *Journal of Empirical Finance*, 1, 83-106.

[27] Engle, R.F. (1982), "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation", *Econometrica*, 50, 987-1008.

[28] Engle, R.F. and G. Gallo (2003), "A Multiple Indicator Model for Volatility Using Intra Daily Data", Discussion Paper NYU and Universitá di Firenze.

[29] Forsberg, L. and E. Ghysels (2004), "Jumps or No Jumps, Smooth or Rough: It Does Not Matter For Predicting Volatility - Choosing the Right Regressor Does," Discussion paper UNC.

[30] Gallant, A. R., C.-T. Hsu, and Tauchen, G. (1999), "Using Daily Range Data to Calibrate Volatility Diffusions and Extract the Forward Integrated Volatility", *Review of Economics and Statistics*, 84, 617-631.

[31] Ghysels, E., P. Santa-Clara and R. Valkanov (2002a), "There is a Risk-return Tradeoff after all," *Journal of Financial Economics* (forthcoming).

[32] Ghysels, E., P. Santa-Clara and R. Valkanov (2002b), "The MIDAS Touch: Mixed Data Sampling Regression," Discussion Paper UCLA and UNC available at: http://www.unc.edu/ eghysels

[33] Ghysels, E., P. Santa-Clara, A. Sinko and R. Valkanov (2004), "MIDAS Regressions: Results and New Directions" Work in progress.

[34] Giraitis, L., P. Kokoszka and R. Leipus (2000), "Stationary ARCH models: dependence structure and Central Limit Theorem", *Econometric Theory*, 16, 3-22, 2000

[35] Granger, C.W.J. (1969), Investigating Causal Relations by Econometric Models and Cross Spectral Methods, *Econometrica*, 37, 428-438.

[36] Greene, W. H. (2000), *Econometic Analysis*, Fourth Ed., Prentice Hall.

[37] Hansen, L.P. and T.J. Sargent (1993) ,"Seasonality and approximation errors in rational expectations models", *Journal Of Econometrics* 55, 21-55.

[38] Hansen, P. and A. Lunde, (2004), "Realized Variance and IID Market Microstructure Noise" Discussion Paper.

[39] Harvey, A.C. (1998) "Long Memory in Stochastic Volatility," in J. Knight and S. Satchell (eds.) *Forecasting Volatility in Financial Markets*, Butterworth-Heinemann, Oxford.

[40] Judge,G., W.E.Griffith, R.C. Hill, H. Lutkepohl and T.-C. Lee (1985) *The theory and Practice of Econometrics*, Second Edition, Jhon Wiley & Sons.

[41] Martens, M., Y.-C. Chang and S. J. Taylor (2002) ,"A Comparison of Seasonal Adjustment Methods when Forecasting Intraday Volatility", *Journal of Financial Research,* 25, 283299.

[42] Merton, R. C. (1980), "On estimating the expected return on the market: An exploratory investigation", *Journal of Financial Economics*, 8, 323-361.

[43] Nelson, D.B. and C.Q. Cao (1992), " Inequality Constraints in the Univariate GARCH Model" *Journal of Business and Economic Statistics*, 10, 229-235.

[44] Sims, C. A. (1974), "Distributed lags", in Intrilligator, M. D. and D. A. Kendrick, *Frontiers of Quantitative Economics II*, North-Holland, Amsterdam.

[45] Sims, C.A. (1993), "Rational expectations modeling with seasonally adjusted data", *Journal Of Econometrics*, 55, 9-19.

[46] Stock, J.H. and M.W. Watson (2003), *Introduction to Econometrics*, Addison-Wesley.

[47] Taylor, S. J. and X. Xu (1997), "The incremental volatility information in one million foreign exchange quotations", *Journal of Empirical Finance*, 4, 317-340.

[48] West, K. (2004), "Forecast Evaluation" in Elliott, G., C.W.J. Granger and A. Timmerman (eds.) *Handbook of Economic Forecasting* (Elsevier, forthcoming).

[49] Woerner J. (2002), "Variational sums and power variation: a unifying approach to model selection and estimation in semimartingale models", Discussion Paper, Oxford University.

[50] Wooldridge, J.M. (1999), *Introductory Econometrics: A Modern Approach*, South-Western.

[51] Wood, R.A., T. H. McInish and J. K. Ord (1985) "An Investigation of Transactions Data for NYSE Stocks", *Journal of Finance*, 40, 723-739.

[52] Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2003), "A tale of two time scales: Determining integrated volatility with noisy high-frequency data", Working Paper w10111, NBER.

Figure 1: Unconstrained Weights on Lagged Daily Predictors

This figure plots the weights (normalized to sum up to one) obtained from an OLS log volatility regression for a four week ($4wks$) prediction horizon with 10 lags of the following daily regressors: (1) realized volatility, denoted by $QV\,MIDAS$; (2) realized power variation, denoted by $PV\,MIDAS$; (3) absolute returns; and (4) daily range, denoted by daily R. The weights are estimated with no restrictions using the entire sample.
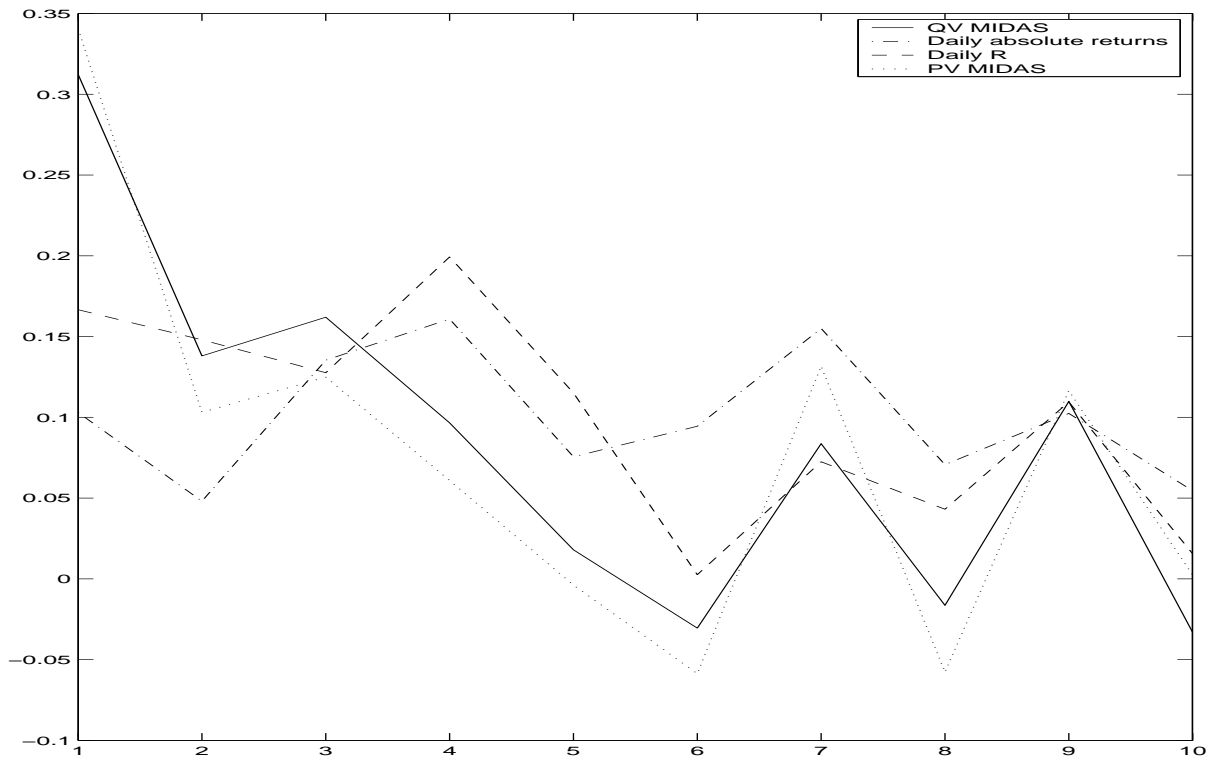
Figure 2: MIDAS and ABDL Log Volatility prediction Four Week Horizon

This figure plots the weights of the MIDAS estimators with a four-week prediction horizon for different regressors: (1) a MIDAS model based on daily squared returns appearing in regression (1.9); (2) MIDAS regression involving past daily realized volatility appearing in regression (1.1) denoted by QV; (3) daily realized power to run regression (1.7) denoted by PV; (4) daily absolute returns to run regression (1.5) denoted ABS and finally; and (5) the daily range for regression (1.6) denoted by R. The solid line is the benchmark ABDL ARFI(5,d) model specification appearing in equation (1.10). All estimates are obtained using the entire sample.
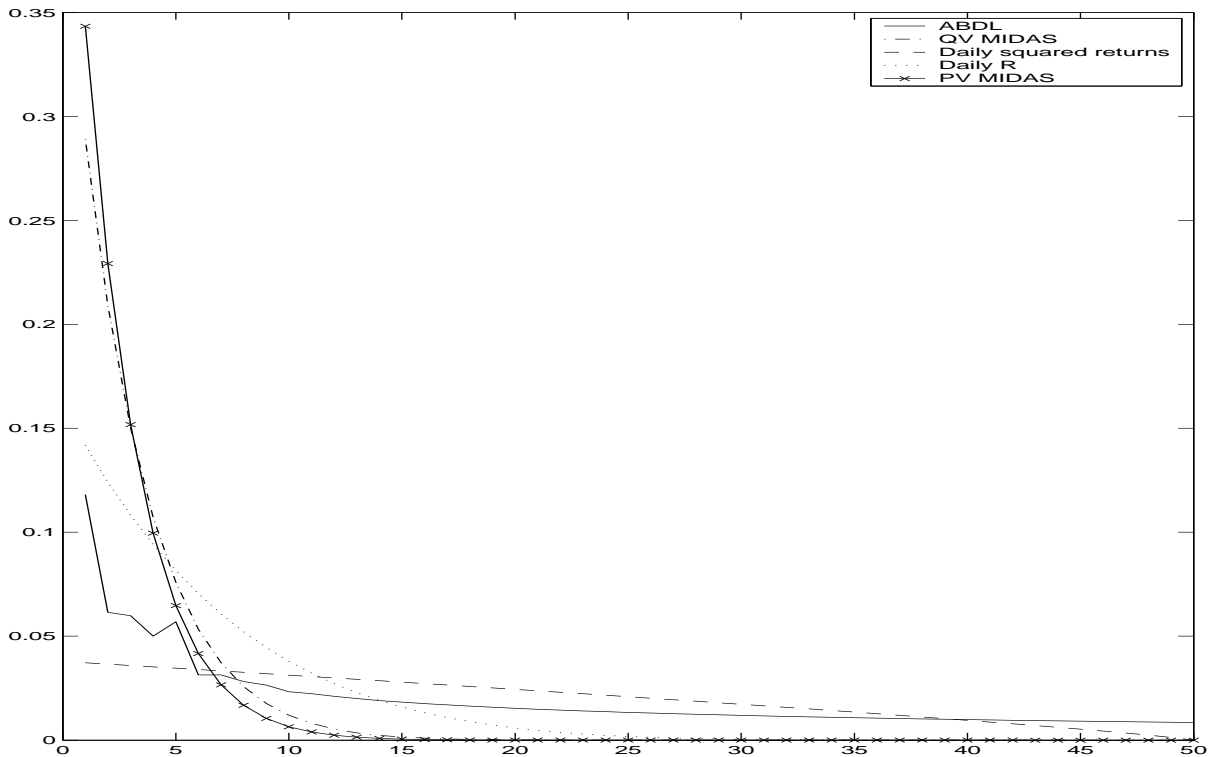
Figure 3: MIDAS Volatility Four Weeks Horizon and Daily Regressors

This figure plots the weights that the MIDAS estimators with a four-weeks ($4wks$) prediction horizon for different regressors: (1) a MIDAS model based on daily squared returns appearing in regression (1.9); (2) MIDAS regression involving past daily realized volatility appearing in regression (1.1) denoted by QV; (3) daily realized power to run regression (1.7) denoted by PV; (4) daily absolute returns to run regression (1.5) denoted by ABS; and (5) the daily range for regression (1.6) denoted by R. All estimates are obtained using the entire sample.
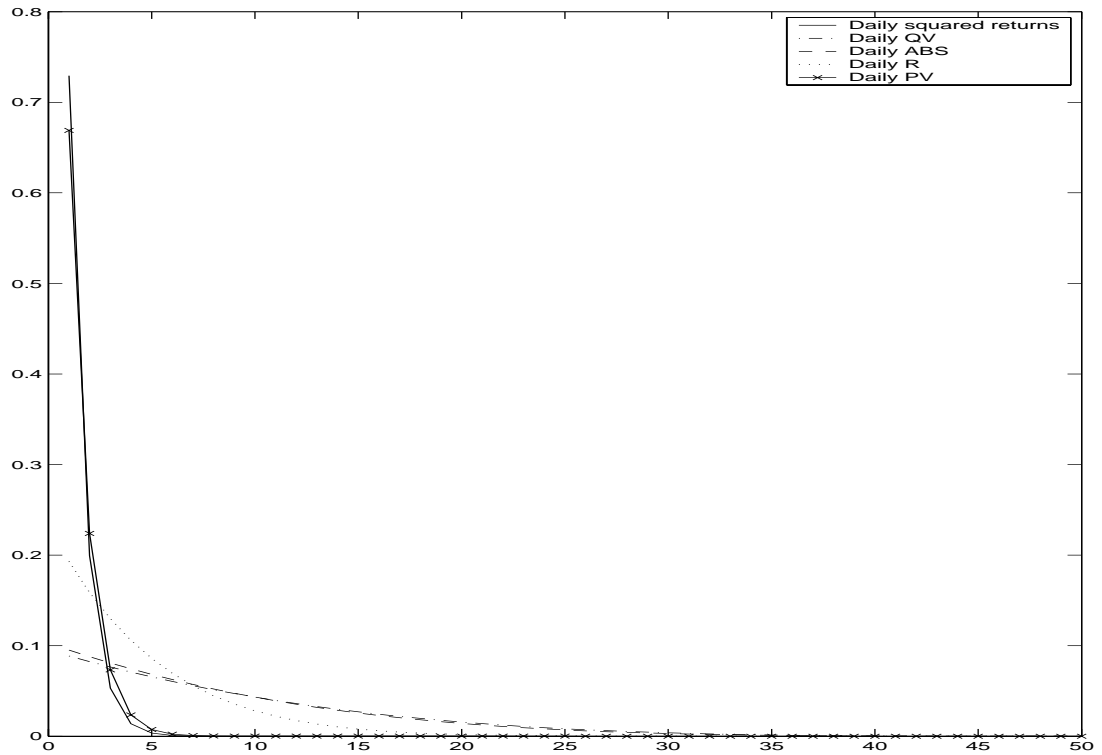
Figure 4: MIDAS Weights One Week Horizon and Intra-Daily Regressors

This figure plots the weights of the MIDAS estimators for one week ($1wk$) prediction horizon for predicting log increments of quadratic variation based on log 5-minute high-frequency intra-daily squared (absolute) returns appearing in regression (3.1). The MIDAS polynomial weights for both hourly adjusted and unadjusted squared returns are displayed. All estimates are obtained using the entire sample.
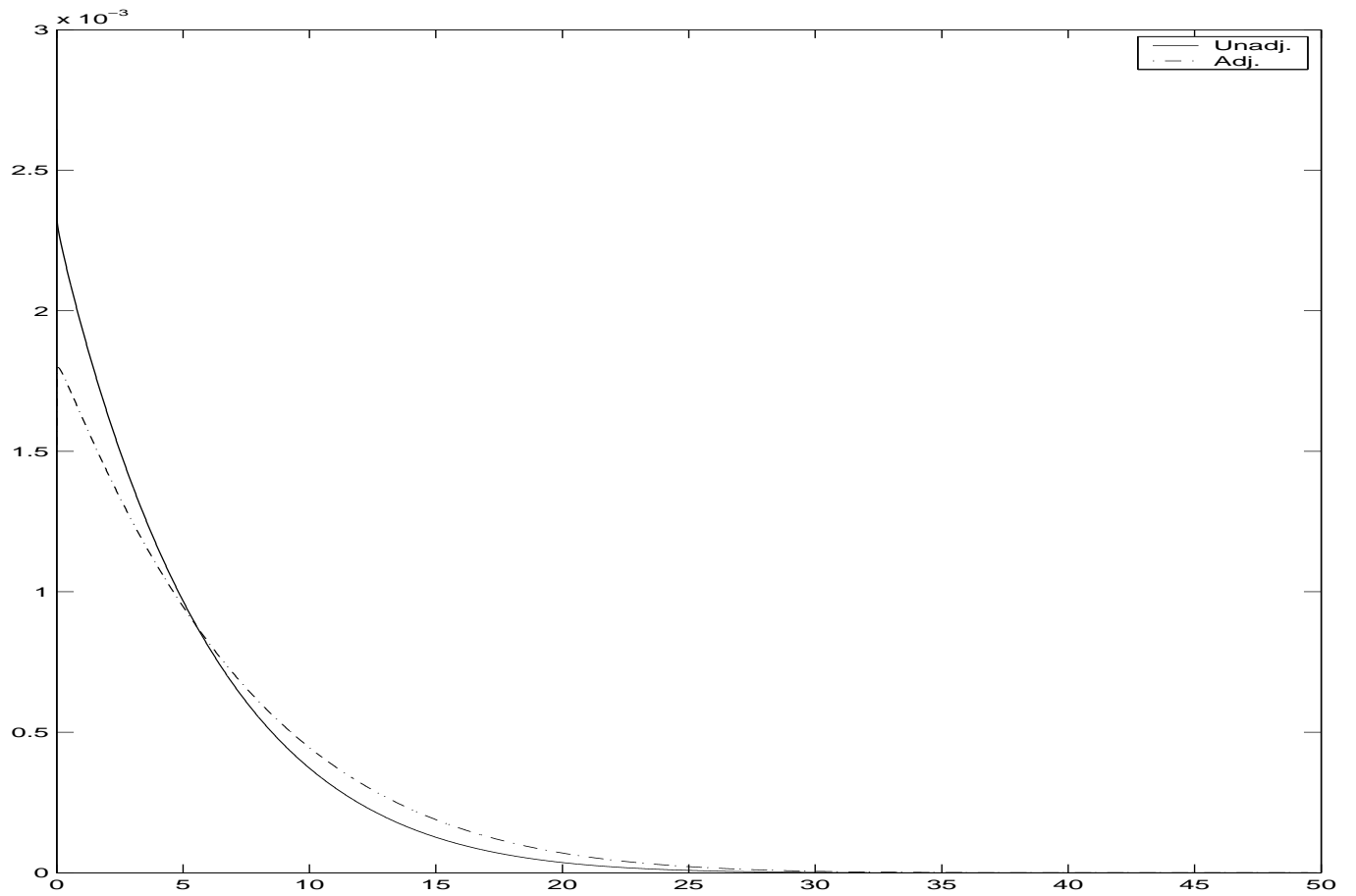
Table 1: In-Sample MSE Comparisons of MIDAS Models with Daily Regressors - DJ Index

Each entry in the table corresponds to the ratios MSE^{MIDAS}/MSE^{ABDL}, where MSE^{ABDL} is the in-sample MSE for the benchmark ARFI(5,d) model in equation (1.10) and MSE^{MIDAS} is the in-sample MSE from the MIDAS model using lagged daily predictors, shown in the corresponding column and discussed in section 1.1. The empirical results are based on a dataset consisting of five-minute intra-day returns of the Dow Jones Composite Portfolio (DJ) over a ten year period, from April 1, 1993 to October 31, 2003 and for a subsample April 1, 1993 to March 31, 2001 (left). The MIDAS regressions are run on a weekly, bi-weekly, tri-weekly, and monthly data sampling schemes with non-overlapping $\tilde{Q}_{t+H,t}^{(Hm)}$, for $H = 5$ days ($1wk$), $H = 10$ ($2wks$), 15 ($3wks$) and 20 ($4wks$) days using two measures of volatility $Q_{t+H,t}$ and $\log(Q_{t+H,t})$, presented in Panels A and B, respectively. The MIDAS regressions are obtained with the Beta lag structure (1.2).

|	Sample Apr. 1, 1993 - Mar. 31, 2001					Sample Apr. 1, 1993 - Dec. 31, 2003				
	$\tilde{Q}_{t,t-1}^{(m)}$	$r_{t,t-1}^2$	$|r_{t,t-1}|$	$[hi-lo]_{t,t-1}$	$\tilde{P}_{t,t-1}^{(m)}$	$\tilde{Q}_{t,t-1}^{(m)}$	$r_{t,t-1}^2$	$|r_{t,t-1}|$	$[hi-lo]_{t,t-1}$	$\tilde{P}_{t,t-1}^{(m)}$
Panel A: $\tilde{Q}_{t,t-H}^{(m)}$ MIDAS with daily lags of regressors										
1 wk	0.871	1.043	0.969	0.856	0.825	0.819	1.023	1.006	0.866	0.788
2 wks	0.851	0.962	0.875	0.791	0.758	0.831	1.118	1.032	0.933	0.711
3 wks	0.793	0.881	0.774	0.722	0.691	0.683	1.016	0.933	0.819	0.649
4 wks	0.805	0.908	0.806	0.743	0.681	0.696	1.043	0.987	0.837	0.606
Panel B: $\log(\tilde{Q}_{t,t-H}^{(m)})$ MIDAS with daily lags of regressors										
	$\log(\tilde{Q}_{t,t-1}^{(m)})$	$\log(r_{t,t-1})^2$	$\log|r_{t,t-1}|$	$\log[hi-lo]_{t,t-1}$	$\log(\tilde{P}_{t,t-1}^{(m)})$	$\log(\tilde{Q}_{t,t-1}^{(m)})$	$\log(r_{t,t-1})^2$	$\log|r_{t,t-1}|$	$\log[hi-lo]_{t,t-1}$	$\log(\tilde{P}_{t,t-1}^{(m)})$
1 wk	0.924	1.509	—	1.062	0.921	0.938	1.574	—	1.066	0.904
2 wks	0.881	1.396	—	1.015	0.867	0.935	1.465	—	1.082	0.896
3 wks	0.821	1.134	—	0.893	0.772	0.902	1.311	—	1.024	0.862
4 wks	0.812	1.177	—	0.921	0.761	0.873	1.326	—	1.060	0.825

Table 2: Regression Diagnostics and Estimated Weights of MIDAS Models with Daily Regressors - DJ Index

The table contains estimated MIDAS regression parameters (equations (1.1) and (1.2)). We present a summary of the weight function $b_H(k,\theta)$ rather than the parameters θ because this is the object of interest. Column "Day 1" represents how much weight is placed on the first daily lag of the predictor, Column "Days 2-5" report how much of the weight is on the second to the fifth lag and so on. The column "Q(10)" reports the Portmanteau test of serial correlation in the residuals. The empirical results are based on a data set consisting of five-minute intra-day returns of the Dow Jones Composite Portfolio (DJ), April 1, 1993 to March 31, 2001. The regressions are run on a weekly, bi-weekly, tri-weekly, and monthly data sampling schemes with non-overlapping $\tilde{Q}^{(Hm)}_{t+H,t}$, for $H = 5$ days ($1wk$), $H = 10$ ($2wks$), 15 ($3wks$) and 20 ($4wks$) days, respectively. The top two panels display the estimates from the forecasting regressions of $Q_{t+H,t}$, while the bottom two panels display the estimates from the forecasting regressions of $\log(Q_{t+H,t})$. The left and right panels are produced using lagged daily realized volatility or lagged daily realized power as predictors. Results from using other predictors (daily squared returns, absolute returns, and ranges) are very similar and are hence omitted for brevity. They are available from the authors.

Sample Apr. 1, 1993 - Mar. 31, 2001

	μ_H	ϕ_H	Day 1	Days 2-5	Days 6-20	> 20 Days	Q(10)	μ_H	ϕ_H	Day 1	Days 2-5	Days 6-20	> 20 Days	Q(10)
	$\tilde{Q}^{(m)}_{t,t-H}$ MIDAS with daily lags of regressors $\tilde{Q}^{(m)}_{t,t-1}$							$\tilde{Q}^{(m)}_{t,t-H}$ MIDAS with daily lags of regressors $\tilde{P}^{(m)}_{t,t-1}$						
1 wk	0.798	4.079	0.155	0.429	0.402	0.014	0.001	-3.250	609.918	0.164	0.443	0.382	0.011	0.009
2 wks	1.987	7.692	0.118	0.363	0.478	0.041	0.191	-5.290	1115.270	0.175	0.459	0.358	0.008	0.060
3 wks	3.281	11.387	0.085	0.286	0.524	0.105	0.738	-8.123	1705.793	0.103	0.330	0.504	0.063	0.700
4 wks	5.675	13.403	0.086	0.288	0.523	0.103	0.329	-8.620	2080.555	0.089	0.296	0.521	0.094	0.719
	$\log(\tilde{Q}^{(m)}_{t,t-H})$ MIDAS with daily lags of regressors $\log(\tilde{Q}^{(m)}_{t,t-1})$							$\log(\tilde{Q}^{(m)}_{t,t-H})$ MIDAS with daily lags of regressors $\log(\tilde{P}^{(m)}_{t,t-1})$						
1 wk	1.657	0.911	0.203	0.491	0.304	0.003	0.102	9.671	1.883	0.129	0.385	0.456	0.030	0.025
2 wks	2.359	0.884	0.194	0.482	0.320	0.004	0.062	10.153	1.832	0.122	0.372	0.470	0.036	0.077
3 wks	2.786	0.881	0.111	0.348	0.490	0.050	0.066	10.415	1.795	0.101	0.325	0.507	0.067	0.161
4 wks	3.074	0.859	0.084	0.283	0.525	0.109	0.370	10.640	1.779	0.061	0.219	0.517	0.202	0.229

Table 3: Out-of-Sample MSE Comparisons of MIDAS Models with Daily Regressors - DJ Index

Each entry in the table corresponds to the ratios MSE^{MIDAS}/MSE^{ABDL}, where MSE^{ABDL} is the out-of-sample MSE for the benchmark ARFI(5,d) model in equation (1.10) and MSE^{MIDAS} is the out-of-sample MSE from the MIDAS model (equations (1.1) and (1.2)) using lagged daily predictors, shown in the corresponding column and discussed in section 1.1. We obtain the out-of-sample forecasts by estimating the models with data from April 1, 1993 to March 31, 2001 and then use the estimates from these regressions (shown in Table 2 above) to forecast future realized volatilities. The regressions are run on a weekly, bi-weekly, tri-weekly, and monthly data sampling schemes with non-overlapping $\tilde{Q}_{t+H,t}^{(Hm)}$, for $H = 5$ days ($1wk$), $H = 10$ ($2wks$), 15 ($3wks$) and 20 ($4wks$) days using two measures of volatility $Q_{t+H,t}$ and $\log(Q_{t+H,t})$. For the log specification, models with $\log r_{t,t-1}^2$ and $\log|r_{t,t-1}|$ regressors are identical, hence the vertical bar to avoid duplication.

	$\tilde{Q}_{t,t-H}^{(m)}$ MIDAS					$\log(\tilde{Q}_{t,t-H}^{(m)})$ MIDAS				
	$\tilde{Q}_{t,t-1}^{(m)}$	$r_{t,t-1}^2$	$|r_{t,t-1}|$	$[hi-lo]_{t,t-1}$	$\tilde{P}_{t,t-1}^{(m)}$	$\log\tilde{Q}_{t,t-1}^{(m)}$	$\log r_{t,t-1}^2$	$\log|r_{t,t-1}|$	$\log[hi-lo]_{t,t-1}$	$\log\tilde{P}_{t,t-1}^{(m)}$
1 wk	0.802	1.048	1.005	0.819	**0.778**	0.746	1.457	—	0.815	**0.729**
2 wks	0.920	1.168	1.039	0.855	**0.726**	0.794	1.296	—	0.882	**0.731**
3 wks	0.814	0.995	0.874	0.794	**0.724**	0.751	1.089	—	0.797	**0.714**
4 wks	0.872	1.089	0.985	0.857	**0.774**	0.893	1.162	—	0.981	**0.854**

Table 4: Out-of-Sample MSE Comparisons of MIDAS Models with Daily Regressors - Individual Stocks

This table displays similar results as in Table 3, but for six individual stocks rather than the DJ Index. More specifically, for each of the six stocks – Disney Co (DIS), General Electric Co (GE), J.P. Morgan Chase (JPM), Exxon Mobil (XOM), MacDonald's (MCD), and Honeywell Co (HON) – each entry in the table corresponds to the ratios MSE^{MIDAS}/MSE^{ABDL}, where MSE^{ABDL} is the out-of-sample MSE for the benchmark ARFI(5,d) model in equation (1.10) and MSE^{MIDAS} is the out-of-sample MSE from the MIDAS model (equations (1.1) and (1.2)) using lagged daily predictors, shown in the corresponding column and discussed in section 1.1. We obtain the out-of-sample forecasts by estimating the models with data from April 1, 1993 to March 31, 2001 and then use the estimates from these regressions (not shown for brevity) to forecast future realized volatilities. The regressions are run on a weekly, bi-weekly, tri-weekly, and monthly data sampling schemes with non-overlapping $\tilde{Q}^{(Hm)}_{t+H,t}$, for $H = 5$ days ($1wk$), $H = 10$ ($2wks$), 15 ($3wks$) and 20 ($4wks$) days using two measures of volatility $Q_{t+H,t}$ and $\log(Q_{t+H,t})$. For the log specification, models with $\log r^2_{t,t-1}$ and $\log|r_{t,t-1}|$ regressors are identical, hence the vertical bar to avoid duplication.

	$\tilde{Q}^{(m)}_{t,t-H}$ MIDAS					$\log(\tilde{Q}^{(m)}_{t,t-H})$ MIDAS				
	$\tilde{Q}^{(m)}_{t,t-1}$	$r^2_{t,t-1}$	$|r_{t,t-1}|$	$[hi-lo]_{t,t-1}$	$\tilde{P}^{(m)}_{t,t-1}$	$\log\tilde{Q}^{(m)}_{t,t-1}$	$\log r^2_{t,t-1}$	$\log|r_{t,t-1}|$	$\log[hi-lo]_{t,t-1}$	$\log\tilde{P}^{(m)}_{t,t-1}$
					Individual stock: DIS					
1 wk	0.895	1.261	1.175	1.027	**0.812**	**0.926**	2.090	—	1.155	0.936
2 wks	0.852	1.149	1.015	0.865	**0.698**	0.806	1.770	—	1.013	**0.804**
3 wks	0.808	1.131	1.000	0.890	**0.684**	0.794	1.826	—	0.984	**0.750**
4 wks	0.761	1.024	0.889	0.824	**0.612**	**0.718**	1.474	—	0.910	0.752
					Individual stock: GE					
1 wk	0.976	1.402	1.336	1.126	**0.839**	0.920	2.046	—	1.219	**0.853**
2 wks	0.843	1.163	1.055	0.980	**0.751**	0.814	1.591	—	0.976	**0.749**
3 wks	0.766	1.015	0.937	0.854	**0.672**	0.742	1.633	—	0.956	**0.692**
4 wks	0.961	0.987	0.829	0.867	**0.807**	0.926	1.306	—	**0.878**	0.882
					Individual stock: JPM					
1 wk	1.042	1.392	1.165	0.923	**0.835**	0.903	1.831	—	1.083	**0.872**
2 wks	0.827	1.281	1.095	0.785	**0.688**	0.814	1.597	—	0.938	**0.780**
3 wks	1.044	1.232	0.995	0.741	**0.727**	0.845	1.487	—	0.863	**0.770**
4 wks	0.932	1.062	0.972	0.722	**0.692**	0.806	1.378	—	0.805	**0.753**

Table continued on next page ...

For a description, please see previous page.

| | $\tilde{Q}_{t,t-H}^{(m)}$ MIDAS | | | | | $\log(\tilde{Q}_{t,t-H}^{(m)})$ MIDAS | | | | |
	$\tilde{Q}_{t,t-1}^{(m)}$	$r_{t,t-1}^2$	$|r_{t,t-1}|$	$[hi-lo]_{t,t-1}$	$\tilde{P}_{t,t-1}^{(m)}$	$\log \tilde{Q}_{t,t-1}^{(m)}$	$\log r_{t,t-1}^2$	$\log |r_{t,t-1}|$	$\log[hi-lo]_{t,t-1}$	$\log \tilde{P}_{t,t-1}^{(m)}$
					Individual stock: XOM					
1 wk	1.005	1.469	1.435	1.066	**0.723**	**0.822**	2.350	—	1.319	0.852
2 wks	0.951	1.254	1.210	0.979	**0.730**	0.804	1.958	—	1.114	**0.791**
3 wks	0.854	1.169	1.121	0.922	**0.767**	0.848	1.545	—	1.002	**0.792**
4 wks	0.949	1.328	1.107	0.954	**0.793**	0.871	1.366	—	0.976	**0.809**
					Individual stock: MCD					
1 wk	1.044	1.394	1.497	1.380	**0.934**	0.960	1.740	—	1.346	**0.922**
2 wks	0.952	1.365	1.499	1.352	**0.834**	0.889	1.740	—	1.336	**0.815**
3 wks	1.143	1.271	1.443	1.286	**0.893**	0.881	1.775	—	1.292	**0.848**
4 wks	0.955	1.082	1.236	1.129	**0.822**	0.894	1.462	—	1.108	**0.833**
					Individual stock: HON					
1 wk	1.087	1.255	1.305	1.019	**0.833**	0.991	1.310	—	1.157	**0.906**
2 wks	0.942	1.128	1.170	0.935	**0.796**	0.925	1.072	—	1.027	**0.842**
3 wks	0.732	1.119	1.141	1.038	**0.552**	0.841	1.072	—	1.169	**0.755**
4 wks	1.160	1.146	1.126	1.050	**0.900**	1.042	**0.861**	—	1.096	0.944

Table 5: In-Sample MSE Comparisons of MIDAS Models with High-Frequency Regressors - DJ Index

The empirical results are based on a data set consisting of five-minute intra-day returns of the Dow Jones Composite Portfolio (DJ) over a ten year period, from April 1, 1993 to October 31, 2003. In conducting the out-of-sample forecasts, we split the sample into two subsamples, April 1, 1993 to March 31, 2001 and April 1, 2001 to October 31, 2003. Therefore, the left panel covers the entire sample, the right one the first subsample. The regressions are run on a weekly, bi-weekly, tri-weekly, and monthly data sampling schemes with non-overlapping $\tilde{Q}_{t+H,t}^{(Hm)}$, for $H = 5$ days $(1wk)$, $H = 10$ $(2wks)$, 15 $(3wks)$ and 20 $(4wks)$ days respectively. The MIDAS regressions are obtained with the Beta lag structure (1.2), imposing $\theta_1 = 1$. We presents the in-sample ratios MSE^{MIDAS}/MSE^{ABDL}, where MSE^{ABDL} is the MSE for the benchmark ARFI(5,d) model appearing in equation (1.10). The results from forecasting the two measures of volatility $Q_{t+H,t}$ and $\log(Q_{t+H,t})$ are presented in Panels A and B, respectively. For the log specification, models with $\log r_{t,t-1}^2$ and $\log |r_{t,t-1}|$ regressors are identical, hence the vertical bar to avoid duplication.

	Unadjusted Model				Hourly Adjusted Model			
	Apr. 1, 1993 - Mar. 31, 2001		Apr. 1, 1993 - Dec. 31, 2003		Apr. 1, 1993 - Mar. 31, 2001		Apr. 1, 1993 - Dec. 31, 2003	
	$r_{t,t-1/m}^2$	$|r_{t,t-\frac{1}{7}m}|$	$r_{t,t-1/m}^2$	$|r_{t,t-1/m}|$	$r_{t,t-1/m}^2$	$|r_{t,t-1/m}|$	$r_{t,t-1/m}^2$	$|r_{t,t-1/m}|$
			$\tilde{Q}_{t,t-H}^{(m)}$ MIDAS with high frequency lags of regressors					
1 wks	0.868	0.825	0.802	0.791	0.903	0.853	0.832	0.831
2 wks	0.849	0.759	0.831	0.709	0.885	0.793	0.875	0.756
3 wks	0.770	0.673	0.661	0.634	0.789	0.696	0.677	0.651
4 wks	0.782	0.680	0.646	0.597	0.724	0.691	0.736	0.610
			$\log(\tilde{Q}_{t,t-H}^{(m)})$ MIDAS with high frequency lags of regressors					
	$\log(r_{t,t-1/m}^2)$	$\log|r_{t,t-1/m}|$	$\log(r_{t,t-1/m}^2)$	$\log|r_{t,t-1/m}|$	$\log(r_{t,t-1/m}^2)$	$\log|r_{t,t-1/m}|$	$\log(r_{t,t-1/m}^2)$	$\log|r_{t,t-1/m}|$
1 wks	0.986	—	0.983	—	1.039	—	1.091	—
2 wks	0.911	—	0.951	—	0.953	—	1.030	—
3 wks	0.736	—	0.893	—	0.745	—	0.959	—
4 wks	0.743	—	0.854	—	0.786	—	0.893	—

Table 6: Out-of-Sample MSE Comparisons of MIDAS Models with High-Frequency Regressors - DJ Index

The empirical results are based on a data set consisting of five-minute intra-day returns of the Dow Jones Composite Portfolio (DJ) over a ten year period, from April 1, 1993 to October 31, 2003. In conducting the out-of-sample forecasts, we split the sample into two subsamples, April 1, 1993 to March 31, 2001 and April 1, 2001 to October 31, 2003. We presents the in-sample ratios MSE^{MIDAS}/MSE^{ABDL}, where MSE^{ABDL} is the MSE for the benchmark ARFI(5,d) model appearing in equation (1.10). All models predict the square root of volatility out of sample, similar to Andersen et al. (2003). The MIDAS regressions are reported in Table 2. Entries in bold show the best performing model for each prediction horizon across separate panels. For the log specification, models with $\log r_{t,t-1}^2$ and $\log |r_{t,t-1}|$ regressors are identical, hence the vertical bar to avoid duplication.

	Unadjusted Model				Hourly Adjusted Model			
	$\tilde{Q}_{t,t-H}^{(m)}$		$\log(\tilde{Q}_{t,t-H}^{(m)})$		$\tilde{Q}_{t,t-H}^{(m)}$		$\log(\tilde{Q}_{t,t-H}^{(m)})$	
	$r_{t,t-1/m}^2$	$|r_{t,t-1/m}|$	$\log r_{t,t-1/m}^2$	$\log |r_{t,t-1/m}|$	$r_{t,t-1/m}^2$	$|r_{t,t-1/m}|$	$\log r_{t,t-1/m}^2$	$\log |r_{t,t-1/m}|$
1 wks	0.797	**0.783**	**0.870**	—	0.880	**0.834**	**1.058**	—
2 wks	0.905	**0.727**	**0.855**	—	0.956	**0.809**	**1.005**	—
3 wks	0.987	**0.883**	**0.815**	—	1.000	**0.892**	**0.927**	—
4 wks	1.035	**0.787**	**0.868**	—	1.259	**0.873**	**0.938**	—