Przegląd Statystyczny. Statistical Review, 2022, vol. 69, 2, 1–19 DOI: 10.5604/01.3001.0016.0363
Received: 12.07.2022, revised: 12.09.2022, accepted: 25.09.2022

# Do mixed-data sampling models help forecast liquidity and volatility?

Barbara Będowska-Sójka,[a] Agata Kliber[b]

**Abstract.** This paper aims to contribute to the existing studies on the Granger-causal relationship between volatility and liquidity in the stock market. We examine whether liquidity improves volatility forecasts and whether volatility allows the improvement of liquidity forecasts. The forecasts based on the mixed-data sampling models, MIDAS, are compared to those obtained from models based on daily data. Our results show that volatility and liquidity forecasts from MIDAS models outperform naive forecasts. On the other hand, the application of mixed-data sampling models does not significantly improve the performance of the forecasts of either liquidity or volatility based on a univariate autoregressive model or a vector-autoregressive one. We found that in terms of the forecasting ability, the VAR models and the AR models seem to perform equally well, as the differences in forecasting errors generated by these two types of models are not statistically significant.

**Keywords:** liquidity, volatility, effective spread estimator, MIDAS
**JEL:** G12, G15

## 1. Introduction

Volatility and liquidity of the financial instruments are the core concepts in empirical finance. The first is usually defined as the statistical measure of the dispersion of returns for a given security, while the second is described as the ability to buy or sell an asset immediately at a low cost without affecting the asset's price significantly (Pástor & Stambaugh, 2003). Volatility and liquidity share some common features: both are unobservable, difficult to estimate and time-varying. There is no simple answer to the question what the best proxy for either volatility (Andersen et al., 2007) or liquidity (Díaz & Escribano, 2020) is. Here two approaches are commonly applied: volatility and liquidity measures are based either on data of the same frequency (e.g. daily measures based on daily data) or on data of higher frequency (e.g. daily measures based on intradaily data) (Ahn et al., 2018; Andersen & Bollerslev, 1998). Generally, measures based on higher-frequency data

[a] Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Econometrics, al. Niepodległości 10, 61-875 Poznań,
e-mail: barbara.bedowska-sojka@ue.poznan.pl, ORCID: https://orcid.org/0000-0001-5193-8304.

[b] Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Applied Mathematics, al. Niepodległości 10, 61-875 Poznań,
e-mail: agata.kliber@ue.poznan.pl, ORCID: https://orcid.org/0000-0003-1996-5550.

should be more informative, as the set of information is more comprehensive (Giot, 2005). However, such data are usually expensive and therefore not available for all investors. The time-varying feature was exhaustively examined both in the case of volatility (Faff et al., 2000) and liquidity (Liang & Wei, 2012). As such, these variables are also difficult to predict.

The aim of the paper is to examine two issues. Firstly, we investigate whether information on the past liquidity can improve volatility forecasts, and vice versa – whether data on previous volatility can improve liquidity forecasts. Secondly, we consider the application of mixed-frequency data by comparing the accuracy of forecasts from mixed-data sampling models, MIDAS (Ghysels et al., 2004), to those which use variables in one frequency only. For the latter, we consider vector autoregressive (VAR) models with the other variable as the regressor, and simple autoregressive (AR) models without any additional variables. We examine which method, the one employing mixed-frequency data or the one applying one-frequency data only, generates better results in terms of out-of-sample volatility and liquidity forecasts.

We employed a dataset from the European emerging market, the Warsaw Stock Exchange (further: the WSE), which was a sample of 118 stocks listed on this market and observed over a period of eight years. Such a forecasting exercise requires liquidity and volatility measures that could be obtained for a low (daily) and high (intradaily) frequencies. Thus, volatility in our approach was approximated by a realised variance (Andersen et al. 2006), while liquidity was calculated as the quoted effective spread of Chung and Zhang (2014). The former measure is identified in the literature as a good proxy for volatility (Andersen & Bollerslev, 1998), and the latter for liquidity (Fong et al., 2017; Ma et al., 2018).

The main result of the study was finding no advantage in using MIDAS models. Models based on daily data only, such as univariate AR or bi-variate VAR ones, performed better than the more complicated AR_MIDAS ones, where mixed frequencies were applied. There was no distinction between the AR and the VAR models – both were performing equally well within the forecasting framework. Among the specifications considered, the MIDAS model outperformed only the naive approach.

The remaining part of the paper is organised in the following way: Section 2 is devoted to the literature review on the dependency between volatility and liquidity, Section 3 describes the sample and variables used in the study, Section 4 shows the research methodology, Section 5 presents empirical results, and Section 6 summarises and concludes the study.

## 2. Literature review

The literature shows that volatility and liquidity are interrelated. Chordia et al. (2001) found that aggregated liquidity is influenced by recent market volatility, among other factors. Rösch and Kaserer (2014) showed that liquidity increases in the time of market downturns, while Yeyati et al. (2008) described the 'spiralling fall' effect, which manifests itself in lower liquidity when stock market returns decrease rapidly and volatility is higher. The faster the market falls, the less liquidity there is. Brunnermeier and Pedersen (2009) showed that higher volatility tends to increase illiquidity, because financial intermediaries reduce their activity in volatile times. Ma et al. (2018) found the dependence between stock market volatility and trading activity, namely as the market becomes more volatile, the trading volume decreases.

Another current in the literature focuses on the causal relationships between liquidity and volatility. There is evidence for a one-direction or bi-directional causality in different stock markets (Będowska-Sójka & Kliber, 2019; Hautsch & Jeleskovic, 2008; Hiemstra & Jones, 1994; Gold et al., 2017). According to the causality definition, if one time series is a Granger cause for another, it improves the latter's forecasts (Ong, 2015). Therefore it seems that combining volatility and liquidity in the forecasting framework and using one of them when predicting the other might be effective.

This study is the extension of the previous research by Będowska-Sójka and Kliber (2019). That former research also pertained to the WSE and showed that there was a causal relationship between volatility and liquidity and vice versa. Moreover, it was demonstrated that liquidity reacted differently to the increase and the decrease in volatility, and likewise volatility – it was affected to a different extent by the rise and the decline in liquidity. A natural extension of that study would be to find out whether the causal relationships are strong enough to be useful in forecasting. Moreover, intraday data seems to be more relevant, as it brings more information about the market than the daily data. Here a question arises whether additional information is useful in predicting the aforementioned measures.

As already mentioned, to address the above issues, we first estimated and then generated volatility and liquidity forecasts from the following models: the MIDAS, the VAR and the AR. In the literature, the MIDAS model is successfully used to describe the dynamics of macroeconomic variables. For instance, Smith (2016) and Maas (2019) used the MIDAS model to successfully nowcast the unemployment by means of Google-search data as a high-frequency regressor. There is evidence that the MIDAS regression outperforms other models in predicting GDP (Ferrara & Marsilli, 2013; Kim & Swanson, 2018; Tsui et al., 2018). Also Andreou et al. (2010) found that using regressions with differently-sampled data improved the forecasting

ability of the empirical economic growth. Other authors showed that incorporating mixed-frequency data to inflation modelling had promising results. Breitung and Roling (2015) demonstrated that the commodity price index is a useful predictor of inflation rates 20–30 days ahead, and Monteforte and Moretti (2013) found that the inclusion of daily variables from the financial market in the model of monthly inflation helps to reduce forecast errors.

Many researchers also proved that the MIDAS model could be successfully applied to both modelling and forecasting of the financial-market data. Although in the MIDAS-GARCH approach (Engle et al., 2013), the high-frequency volatility is modelled with low-frequency data, as e.g. economic indicators (Asgharian et al., 2013; Engle et al., 2013) or other regressors of lower frequency (Ma et al., 2019), there have also been attempts to model the daily volatility with intradaily data. Such a mixed-data sampling approach was applied to volatility prediction by Ghysels et al. (2006), who used high- and low-frequency data to forecast volatility. Their model allowed the improvement of forecasts by 30% compared to the benchmark model. Further, Santos and Ziegelmann (2014) juxtaposed several multi-period volatility forecasting models from the MIDAS and the HAR families in order to forecast the future volatility of the BOVESPA index. They concluded that regressors involving volatility measures robust to jumps are better in forecasting the future volatility – which corroborates the findings described in Ghysels et al. (2006) – and that the relative forecasting performances of the three approaches are comparable.

To our best knowledge, there have not been so far any such attempts when liquidity and volatility were forecasted. There is still no evidence whether the incorporation of high-frequency measures of volatility (or liquidity) is helpful when forecasting liquidity (or volatility) in daily frequency. The presence of causality between volatility and liquidity justifies such experiments.

## 3. The description of the dataset used in the study

### 3.1. Data source and sample description

The sample duration extended from January 2009 until December 2016. The stocks included in the sample were constantly listed on the WSE throughout this period. The final sample consisted of 118 stocks well-established in the market and with a relatively long history of listing (the full list is available from the corresponding author upon request). We used high-frequency data from the database constructed on the basis of data offered directly by the WSE.

As the original source were tick-by-tick data, they had to be pre-processed. The procedures described in Barndorff-Nielsen et al. (2009) were applied, which made it

possible to control for outliers, multiple or missing records, and other incidents that might occur in high-frequency datasets. Then the filtered tick-by-tick data were aggregated into equally-sampled 10-minute, 30-minute, and 60-minute data. Thus we received eight years of data for 118 stocks with four different frequencies: three intraday and one daily.

### 3.2. Volatility and liquidity proxies

As volatility and liquidity are unobservable, we used non-parametric measures to calculate the daily and intradaily estimates. The choice of the proxies was based on the fact they were relatively easy to calculate and it was possible to obtain the estimates in different frequencies: daily and intradaily. Volatility was proxied by realised variance (RV), and calculated as (Andersen et al., 2007):

$$RV_t = \sum_{i=1}^{I} r_{i,t}^2, \tag{1}$$

where $RV_t$ is a daily realised variance in day $t$, $r_{it}$ is a log return in interval $i$ (e.g. 10-minute), and $I$ is the number of intra-daily periods within a day. The realised variance is one of the estimators of volatility that are most frequently used (Andersen & Bollerslev, 1998; Fuertes & Olmo, 2012; Laurent & Violante, 2012).

Out of various liquidity proxies, we chose the closing quoted spread, CQS, of Chung and Zhang (2014). The following formula was applied:

$$CQS_t = \frac{A_t - B_t}{0.5(A_t + B_t)}, \tag{2}$$

where $B_t$ and $A_t$ were the bid and the ask prices, respectively, at the end of a given day *t*. Various studies showed that the CQS is the best proxy for unobserved liquidity (Chung & Zhang, 2014; Díaz & Escribano, 2020; Fong et al., 2017).

We also calculated these two measures in the high-frequency setting: the realised volatility were the squares of intradaily returns in a given sampling frequency, while the quoted effective spread was calculated on the basis of the last bid and ask price within a given time interval (e.g. 1 hour).

## 4. Methodology

### 4.1. The MIDAS model

We used the following notation: $y_t$ was a dependent variable representing a low--frequency process and sampled at daily frequency while $x_\tau$ was an explanatory variable sampled in high frequency. For $x_\tau$, we considered three distinct cases: a 10-minute, a 30-minute, and a 60-minute frequency. Please note that $By_t = y_{t-1}$ and $Lx_\tau = x_{\tau-1}$ were the lags of the low-frequency and the high-frequency processes, respectively. It was assumed that for each low-frequency period $t = t_0$, we observed high frequency process $x_\tau^{(i)}$ at $m_i \in N$ intervals: $\tau = (t_0 - 1)m_i + j$, $j = 1, \dots, m_i$.

Since the session schedule within our sample period changed three times, we choose to consider records from 9.00 a.m. to 4 p.m. Due to some irregularities in the data, and in order to conveniently define equally sampled series, we had to skip the first observation, when data was sampled at the frequency higher than 1 hour. For 10-minute data, the first observation was made at 9.10 a.m., while the last one was recorded at 4 p.m. Thus we have $m_1 = 42$ observations of the high-frequency process, and $\tau = 0, \dots, 42$. When $x$ was sampled at a 30-minute frequency, the first observation came at 9.30 a.m., while the last one was made at 4 p.m., thus: $m_2 = 14$ and $\tau = 0, \dots, 14$. Finally, when $x$ was sampled every 60 minutes, the first observation came at 9:00 a.m., while the last one was made at 4 p.m., so $m_3 = 8$ and $= 0, \dots, 8$. In each of the above cases, there was only one observation per day for the low-frequency process.

The MIDAS model can be written in a compact form as (Ghysels et al., 2016):

$$\alpha(B)y_t = \beta(L)^T \mathbf{x}_{t,0} + \epsilon_t, \tag{3}$$

where:
$\alpha(z) = 1 - \sum_{j=1}^{p} \alpha_j z^j$ (low-frequency lag operator),
$\mathbf{x}_{t,0} \coloneqq \left(x_{tm_0}^{(0)}, \dots, x_{tm_i}^{(i)}, \dots, x_{tm_l}^{(l)}\right)^T$,
$\beta(z) = 1 - \sum_{j=1}^{p} \beta_j z^j$ (high-frequency lag operator), $\beta_j = \left(\beta_j^{(0)}, \dots \beta_j^{(i)}, \beta_j^{(l)}\right)$,
$T$ denotes transposition, and $i$ the frequency period.

In our study, we considered AR(1)-MIDAS models, and in each model we included explanatory variables of only one frequency (either 10-minute, 30-minute, or 60-minute). Thus, our model can be specified in an alternative form as:

$$y_t = ay_{t-1} + \sum_{j=0}^{l} \beta_j x_{tm-j} + \epsilon_t\,. \tag{4}$$

To estimate the model, one needs to align the high-frequency data to the low-frequency data. The alignment is performed through the following transformation (Ghysels et al., 2016):

$$\sum_{j=0}^{l} \beta_j x_{tm-j} = \sum_{r=0}^{q} \lambda_r \tilde{x}_{t-r}\,, \tag{5}$$

where $q \in N$ denotes a low-frequency number of lags, and $\tilde{x}_{t-r}$ the parameter--driven low-frequency aggregates (Ghysels et al., 2016):

$$\tilde{x}_{t-r} = x_{t-r}(\boldsymbol{\delta}_r) = \sum_{s=1}^{m} \omega_r(\boldsymbol{\delta}_r; s) x_{(t-r-1)m+s}\,, \tag{6}$$

The function $\omega_r(\boldsymbol{\delta}_r; s)$ is called a weighting function, and its parameter vector $\boldsymbol{\delta}_r$ can generally vary with each variable and low-frequency lag order $r$. The aggregation weights $\lambda_r$ are usually non-negative and satisfy the normalisation constraints: $\sum_{s=0}^{m-1} \omega_r(\boldsymbol{\delta}_r; s) = 1$. To have the weights add to one, it is convenient to define a weighting function in the following form (Ghysels et al., 2016):

$$\forall r\colon \omega_r(\boldsymbol{\delta}_r; s) = \frac{\psi_r \omega_r(\boldsymbol{\delta}_r; s)}{\sum_{j=1}^{m} \psi_r(\boldsymbol{\delta}_r; j)}, \qquad s = 1, \ldots, m, \tag{7}$$

where $\psi_r(\cdot)$ denotes some underlying function. If the latter is non-negatively valued and the denominator is positive, the weights (7) are also non-negative (Ghysels et al., 2016).

There are various possible specifications of the underlying functions described in the literature: an exponential Almon lag polynomial, beta function, Gomperts, log--Cauchy, etc. (see Ghysels et al., 2016 for details). Using the constraint function has two advantages. Firstly, it allows the reduction of the number of parameters in the model. Secondly, if the parameters of an underlying data-generating process follow a certain functional constraint, and this constraint is well-approximated by a chosen constraint function, the accuracy of the out-of-sample predictions can improve significantly – as shown by Ghysels et al. (2016).

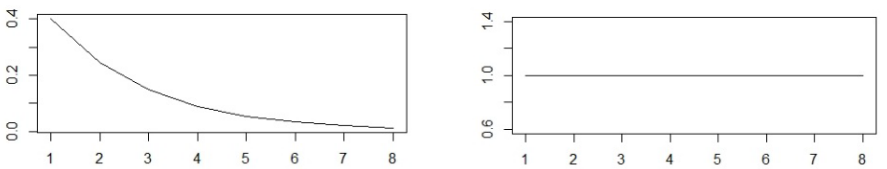
In our study, we use the exponential Almon lag polynomial:

$$\psi(\boldsymbol{\delta}; s) = \exp\left(\sum_{j=1}^{p} \delta_j s^j\right), p \in N, \tag{8}$$

in its normalized and non-normalized form. The Almond polynomial is flexible and can take various shapes with only a few parameters (Ghysels et al., 2007). As a starting point for the estimation, we parametrised the constraint function in such a way that the newest observations had higher weights than the older ones (Almon

function with two parameters: 1 and –0.5). As an alternative, equal weights were tested. We present the lag functions in Figure 1.

**Figure 1.** Alternative shapes of weight functions used in the MIDAS model specification




Source: authors' calculations.

There are several phases of a model selection. In each of them, we took into consideration all the information from the day (i.e. all eight intra-daily observations of the regressor in the hourly model, and 42 observations in the 10-minute model). We chose the optimal constraint function based on the AIC criterion. In order to check the adequacy of functional constraints, we performed the heteroscedasticity and autocorrelation robust weight specification test (hARh) (Ghysels et al., 2016). If a model did not pass the test, we computed the 'unrestricted' MIDAS model, imposing no constraints on the regression parameters (see: Foroni et al., 2011) for the comparison of the unrestricted MIDAS models with the models with the Almond constraints). Next, the forecasts for the chosen model were generated. Our preliminary research demonstrated that the best results were obtained for the AR-MIDAS (not the simple MIDAS), therefore we used it. As the liquidity and volatility measures are non-stationary, we obtained the first differences in the variables. In the estimation of the AR-MIDAS models and the generation of the forecasts, we used the following R packages: midasr (Ghysels et al., 2016), forecast (Hyndman et al., 2019; Hyndman and Khandakar, 2008) and highfrequency (Boudt et al., 2018).

### 4.2. Vector autoregressive model

In the next step, we also computed forecasts of liquidity and volatility using the vector autoregression (further: the VAR). The VAR model has the following form:

$$\begin{cases} \Delta VOL_t = \sum_{i=1}^{k} \alpha_{1i} \Delta VOL_{t-i} + \sum_{i=1}^{k} \beta_{1i} \Delta LIQ_{t-i} + \epsilon_{1t} \\ \Delta LIQ_t = \sum_{i=1}^{k} \alpha_{2i} \Delta VOL_{t-i} + \sum_{i=1}^{k} \beta_{2i} \Delta LIQ_{t-i} + \epsilon_{1t} \end{cases}, \tag{9}$$

where $\Delta VOL_t$ denotes the change of volatility in day $t$, $\Delta LIQ_t$ is the change of liquidity between day $t-1$ and $t$, $\epsilon_t$ is the $iidN(0;\sigma)$ term, and $k \leq 5$. Two R packages were applied: vars (Pfaff, 2008a; 2008b) and VAR.etp (Kim, 2014).

Comparing the forecasts from the VAR with the forecasts from the MIDAS enabled us to check whether it was better to use only daily data on volatility and liquidity, or daily and intradaily data. The lag length was determined on the basis of the AIC criterion with the maximum allowed length of the lag being 5 days. Thus, we assumed that the impact of the information from a period longer than a week was not significant for the prediction purposes. In order to maintain consistency with previous approaches, the length of the out-of-sample interval was 100 days. We generated the one-day-ahead forecasts and computed the mean absolute error, MAE, and root mean square errors, RMSE (Hyndman & Koehler, 2006), according to the following formulas:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_{T+i} - \hat{y}_{T+i}|, \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_{T+i} - \hat{y}_{T+i})^2}, \tag{11}$$

where $y_{T+i}$ denotes the $i$-th value of the out-of-sample dependent variable, $\hat{y}_{T+i}$ its forecast, and $N$ is the number of the out-of-sample length (in our case $N = 100$).

### 4.3. Autoregressive model and the naive approach

Eventually, we investigated whether the presence of regressors in the model improved the forecasts. To verify this, we compared the forecasts from the AR-MIDAS models with those obtained by means of the simple autoregressive models and the naive approach. By the 'simple AR' we mean the models with autoregressive variables in daily frequency only, without the MIDAS part. The lag number in the AR model was determined on the basis of the AIC information criterion, with the restriction on the maximum length of the lag to 5 days (as in the case of the VAR model). The length of the out-of-sample interval was 100 days. We generated the forecasts for one day ahead ($h = 1$), and computed the RMSE and the MAE errors.

## 5. Empirical results

In the empirical part of the paper, we estimated two sets of models – one in terms of the changes in volatility, and the other in terms of the changes in liquidity:

1. The change in the daily volatility approximated by RV is modelled as an autoregressive process with high-frequency explanatory variables involving changes in the 10-, 30- or 60-minute liquidity measured as the CQS:

$$\Delta RV_t = \alpha_i \Delta RV_{t-1} + \sum_{j=0}^{l} \beta_j \Delta LIQ_{m-j} + \epsilon_t. \tag{12}$$

   The alternative models are the VAR model with the CQS daily liquidity measure, a simple AR model for RV, and the naive approach.
2. The change in the daily liquidity represented by the CQS proxy was modelled as an autoregressive process with high-frequency explanatory variables standing for the changes in volatility, proxied by the squared returns calculated in 10-, 30- or 60-minute intervals (SQRET):

$$\Delta LIQ_t = \alpha_i \Delta LIQ_{t-1} + \sum_{j=0}^{l} \beta_j \Delta SQRET_{m-j} + \epsilon_t. \tag{13}$$

The following were the alternatives: the VAR model with RV calculated using data of the same frequency as the CQS, the AR models for the liquidity measure, and the naive approach.

We generated the forecasts from each model for each measure considered and we compared the forecast errors, calculated as the MAE and the RMSE. To assess the forecast ability of the models, in the first step we computed the ratio of the errors from the AR-MIDAS, the VAR and the AR to the naive errors. Additionally, we used the Diebold-Mariano test to verify the hypotheses that the values of the errors produced from the MIDAS model are smaller than the ones produced by the VAR, AR and the naive methods.

### 5.1. Forecasting ability of volatility models

The results of the forecasting study for realised volatility are presented in Table 1. The table shows the relationship between the one-step-ahead forecast errors from the AR-MIDAS model and the alternative approaches, i.e. the VAR model, the AR model and the naive approach. We considered forecasts of the volatility estimates, i.e. the daily RV, calculated on the basis of the 60-, 30- or 10-minute data. The sampling frequencies of the variables used in the models were consistent: the daily RV was based on the same frequency as the explanatory variables used in the models. With regard to the VAR models and all the sampling frequencies, the

out-of-sample relative forecast errors were larger than 1. It means that in all the cases, the forecasts obtained by means of the AR-MIDAS models were less accurate than those received from the VAR models. The higher the sampling frequency, the larger the discrepancy in the predictions observed in the case of the VAR.

**Table 1.** Forecasting ability of the AR-MIDAS method compared to the VAR, AR and naive methods – realised volatility

| Comparison method | Frequency of regressor | Alternative model | | |
|---|---|---|---|---|
| | | VAR | AR | NAIVE |
| RMSE | 60 min | 113.74% | 114.84% | 55.89% |
| | 30 min | 116.42% | 117.31% | 57.27% |
| | 10 min | 123.01% | 123.70% | 61.58% |
| MAE | 60 min | 110.78% | 112.90% | 58.25% |
| | 30 min | 114.63% | 116.58% | 60.20% |
| | 10 min | 120.93% | 122.26% | 63.85% |
| Diebold-Mariano test | 60 min | 1.69% | 0.00% | 98.31% |
| | 30 min | 0.85% | 0.00% | 95.76% |
| | 10 min | 0.00% | 0.00% | 94.92% |

Note. The upper and middle parts of the table present the values of the relative errors of the prognosis calculated as a percentage ratio of the one-ahead forecast error from the AR-MIDAS model for volatility and from one of the alternative models: 1) the VAR model 2) the AR model or 3) the naive approach to the realised volatility value. Two types of errors are provided, i.e. the RMSE and MAE errors. The lower part of the table presents the percentage of cases where the forecast errors from the AR-MIDAS model were more accurate than the forecasts from the VAR model, the AR model or the naive forecasts. This comparison was performed for the MAE error on the basis of the Diebold-Mariano test. The numbers represent the percentage of cases where the AR-MIDAS model proved more effective than any of the alternatives. The results are shown separately for different sampling frequencies: 60-, 30- and 10-minute frequencies.

Source: authors' calculations.

When we compared the AR-MIDAS with a simple AR model without the MIDAS part, the results were similar. The values of forecasting errors from the AR-MIDAS models were definitely higher for the 10-minute data. The forecasts of daily RV-generated on the basis of the AR-MIDAS model were generally less accurate than those based on the simple AR model.

The results were quite opposite, however, when we compared the AR-MIDAS to the naive approach. In all the cases, the relative forecast errors were less accurate in the case of the AR-MIDAS than the naive approach. The forecasts based on the AR-MIDAS model were more accurate for shorter forecast horizons. Also, an improvement was observed in the forecast accuracy when the frequency of the explanatory variable was lower (e.g. 60-minute frequencies were used instead of 10-minute ones). Thus, the AR-MIDAS model proved to have an advantage over the

naive method which, on the other hand, diminished as the VAR or the AR specification was used.

We also considered a different forecast error measure, i.e. the MAE, and examined the robustness of the results (see the middle part of Table 1). The results demonstrated that the VAR model allowed the generation of more accurate forecasts than the AR-MIDAS model. Moreover, no changes were observed in the results for the AR model nor the naive approach. The forecasting ability of the former is always higher than that of the AR-MIDAS specification, while the opposite is true for the latter.

We also applied the Diebold-Mariano test (Diebold, 2015) to compare the predictive accuracy and to verify whether the differences in the forecast errors resulting from the AR-MIDAS and those resulting from the three remaining approaches were significantly different from 0. We used a one-sided test where the null hypothesis stated that there were no differences between the two series of forecast errors, while the alternative hypothesis stated that the forecast errors of the AR-MIDAS model were less significant than those of the VAR model, AR model or the naive method. The test was applied to the forecasts generated separately for each stock and the results were averaged across the sample. The final result showed how often the predictive accuracy of the AR-MIDAS model was higher than that obtained from the remaining models in the cross section.

The lower part of Table 1 shows the results of the Diebold-Mariano test. We found that the AR-MIDAS model is more accurate only when compared with the naive approach. A simple AR model is always more accurate than the AR-MIDAS model, while the VAR model for the same horizon is almost always more accurate than the AR-MIDAS model. We also argue that, based on the results of the Diebold-Mariano test for errors, there is no need to employ a mixed-data sampling model in this particular forecasting case. The sole application of a simple AR or VAR model would generate more accurate forecasts of volatility.

### 5.2. Forecasting ability of liquidity models

In this section of the paper we consider forecasts of liquidity. The upper part of Table 2 shows the relative forecasting RMSE. We found that in terms of liquidity forecasts, the VAR and AR models were always more accurate than the AR-MIDAS model. Similarly to the volatility forecasting, the naive approach generated less accurate liquidity forecasts than those obtained on the basis of the AR-MIDAS model in all the considered frequencies.

Additionally, as in the case of the volatility forecasting, we investigated the relative MAE errors (see the middle part of Table 2). Here the results were slightly different:

in most cases the errors resulting from the application of the VAR model were greater than those resulting from the use of the AR-MIDAS model. The only exceptions were the forecasts for one day in 10-minute frequencies The same results were obtained for simple AR models, where the relative errors were lower than 1%, which indicated a slight predominance of the AR-MIDAS model. As far as the RMSE errors were concerned, the naive approach was still less accurate than the AR-MIDAS model.

We also provide the results of the Diebold-Mariano test. The lower part of Table 2 presents the percentage of cases where forecasts obtained by means of the AR-MIDAS models were of higher accuracy than the forecasts obtained by means of the alternative models. We found that as regards both the VAR and the AR models, in most cases their accuracy was higher than that of the AR-MIDAS. When the naive model was considered, its accuracy was in all cases lower than that of the AR-MIDAS.

**Table 2.** The forecasting ability of the AR-MIDAS compared to the VAR, AR and naive methods: liquidity

| Comparison method | Frequency of regressor | Alternative model | | |
|---|---|---|---|---|
| | | VAR | AR | NAIVE |
| RMSE | 60 min | 105.49% | 105.62% | 58.24% |
| | 30 min | 106.06% | 106.16% | 57.23% |
| | 10 min | 107.89% | 107.98% | 58.38% |
| MAE | 60 min | 96.22% | 96.43% | 57.26% |
| | 30 min | 99.13% | 99.24% | 56.08% |
| | 10 min | 100.42% | 100.59% | 56.54% |
| Diebold-Mariano test | 60 min | 22.88% | 21.19% | 100.00% |
| | 30 min | 11.86% | 11.02% | 100.00% |
| | 10 min | 6.78% | 7.63% | 100.00% |

Note. The upper and middle parts of the table present the values of the relative errors calculated as a percentage ratio of the one-ahead forecast error from the AR-MIDAS model for liquidity and from one of the alternative models, namely the VAR model, the AR model or the naive approach to the realised liquidity value. Two types of errors are provided, i.e. RMSE and MAE errors. The lower part of the table presents the percentage of cases where forecast errors from the AR-MIDAS were of higher accuracy than the forecasts from the VAR model, the AR model or the naive forecasts. This comparison was performed for the MAE error on the basis of the Diebold-Mariano test. The numbers represent the percentage of cases where the AR-MIDAS model was more effective than any of the alternatives. The results are shown separately for 60-, 30- and 10-minute frequencies.

Source: authors' calculations.

It is also worth noting that the percentage of cases where the AR-MIDAS outperformed the VAR or the AR models was higher when liquidity was predicted using intradaily volatility rather than vice versa. This indicates that the changes in intradaily volatility influence the dynamics of daily liquidity more often than the

changes of intradaily liquidity influence daily volatility. This suggests that investors observe the changes of prices during the day and on this basis make decisions as to whether to change their position in the asset. In other words – what influences the decision to change the position is more often the movement of prices rather than the interest of other market participants.

### 5.3. Are liquidity and volatility self-explanatory processes? A comparison of the AR and the VAR models

Research carried out to date shows that for both volatility and liquidity forecasting, the VAR and the AR models generate on average more accurate forecasts than the AR-MIDAS models. On the other hand, the latter are better in terms of forecast accuracy than forecasts generated by means of the naive approach. However, the question as to which out of the two, the VAR or the AR, is more effective in forecasting either volatility or liquidity, remains to be answered.

The 'Volatility' column of Table 3 presents a comparison of the forecast errors, i.e. the relative forecast errors from the volatility forecasts based on VAR and AR models. We found that, regardless of the frequency of the data and the forecast error measure, AR models generate a slightly lower number of errors.

The same approach was applied to liquidity forecasts. The results are presented in the 'Liquidity' column of Table 3. They show that all the fractions are very close to 1%, which means that there is no significant difference between forecasts generated through the VAR or the AR model.

**Table 3.** Forecast error of volatility and liquidity changes: comparison of the VAR and AR model

| Error type | Data frequency | Volatility | Liquidity |
|---|---|---|---|
| RMSE | 10 min | 100.56% | 100.08% |
| | 30 min | 100.79% | 100.07% |
| | 60 min | 101.16% | 100.11% |
| MAE | 10 min | 101.06% | 100.16% |
| | 30 min | 101.81% | 100.10% |
| | 60 min | 102.42% | 100.21% |

Note. The table presents the percentage ratio of the RMSE and MAE forecast errors from the VAR and AR models for different forecast horizons. In the VAR model we take into account the lagged daily volatility (RV) and liquidity (CQS).

Source: authors' calculations.

The above leads to the conclusion that in the case of volatility, the AR model might generate slightly better forecasts in terms of accuracy than the VAR model, while in the case of liquidity, the forecasts from both models are equally accurate.

## 6. Discussion and conclusions

Literature on the undertaken subject provides evidence for one-direction or bi--directional causality between volatility and liquidity. The research presented in this paper aimed to verify whether this dependence could be used to improve forecasts of both volatility and liquidity. Four approaches were considered: the first was based on a mixed sampling of data where daily forecasts of volatility (or liquidity) were generated on the basis of the intraday liquidity (or volatility) measures. The second approach was a VAR model based on daily variables only. The two alternatives – a simple AR model and the naive approach – employed only previous realisations of the processes for which the forecasts were generated.

We have found that although using cross-dependency between volatility and liquidity has its advantages, the employment of mixed-data sampling models is not justified. MIDAS models provide more accurate forecasts than those based on the naive approach. However, in the case of volatility forecasts, both the VAR models with lagged volatility and liquidity and the AR models with lagged liquidity generate errors of lower values than the forecasts based on the MIDAS specifications. With regard to liquidity forecasts, there is no significant difference in forecast accuracy between the MIDAS models and the VAR or AR specification. Thus, the values of forecast errors of volatility are lower when one uses previous values of volatility and previous liquidity data in daily frequency only. However, the computational burden and the associated effort of employing the MIDAS model is much greater than that entailed by the simple AR or VAR model. When only the two latter are compared, the ratio of their respective errors is close to one, indicating that there are only negligible differences between both approaches.

Additionally, the prevalence of the VAR and AR models over the MIDAS model becomes even more evident with the application of higher-frequency data (e.g. 10-minute instead of 60-minute data). Results thus produced are important for investors as well as risk managers who might be wondering if it is worth employing more advanced models that require enormous computing power. Our empirical study shows that in the case of liquidity and volatility forecasting, the gains obtained from the use of MIDAS models are rather negligible. The outcome also sheds some light on the behavioural aspect of investing on the WSE. Considering the fact that the percentage of cases where the AR-MIDAS outperformed the VAR or AR models was higher when liquidity was predicted using intradaily volatility than when

volatility was predicted using intradaily liquidity, the conclusion is that what influences the decision on the change of the position is more often the movement of prices rather than the interest of other market participants. The authors' further research in this area will be devoted to examining the stability of these results by means of other volatility and liquidity measures.

## Acknowledgements

This work was supported by the National Science Centre in Poland under grant no. UMO-2017/25/B/HS4/01546. We would like to thank the seminar participants at Adam Mickiewicz University (SEFIN), as well as the attendants of the INFINITI Conference in 2019 in Glasgow and the 26th International Conference of Forecasting Financial Markets in Venice, for their valuable comments. All errors are our own.

## Data availability statement

The data that support the findings of this study are available from the corresponding author upon request.

## References

Ahn, H.-J., Cai, J., & Yang, C.-W. (2018). Which Liquidity Proxy Measures Liquidity Best in Emerging Markets?. *Economies*, *6*(4), 1–29. https://doi.org/10.3390/economies6040067.

Andersen, T. G., & Bollerslev, T. (1998). Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. *International Economic Review*, *39*(4), 885–905. https://doi.org/10.2307/2527343.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., & Diebold, F. X. (2006). Volatility and Correlation Forecasting. In G. Elliott, C. W. J. Granger, A. Timmermann (Eds.), *Handbook of Economic Forecasting* (vol. 1, pp. 777–878). Elsevier. https://doi.org/10.1016/S1574-0706(05)01015-3.

Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007). Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. *Review of Economics and Statistics*, *89*(4), 701–720. https://doi.org/10.1162/rest.89.4.701.

Andreou, E., Ghysels, E., & Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, *158*(2), 246–261. https://doi.org/10.1016/j.jeconom.2010.01.004.

Asgharian, H., Hou, A. J., & Javed, F. (2013). The importance of the macroeconomic variables in forecasting stock return variance: A GARCH-MIDAS approach. *Journal of Forecasting*, *32*(7), 600–612. https://doi.org/10.1002/for.2256.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2009). Realized kernels in practice: trades and quotes. *Econometrics Journal*, *12*(3), C1–C32. https://doi.org/10.1111/j.1368-423X.2008.00275.x.

Będowska-Sójka, B., & Kliber, A. (2019). The Causality between Liquidity and Volatility in the Polish Stock Market. *Finance Research Letters*, *30*, 110–115. https://doi.org/10.1016/j.frl.2019.04.008.

Boudt, K., Cornelissen, J., Payseur, S., Nguyen, G., & Schermer, M. (2018). *Highfrequency: Tools for Highfrequency Data Analysis. R Package version 0.5.3*.

Breitung, J., & Roling, C. (2015). Forecasting inflation rates using daily data: A nonparametric MIDAS approach. *Journal of Forecasting*, *34*(7), 588–603. https://doi.org/10.1002/for.2361.

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market Liquidity and Funding Liquidity. *Review of Financial Studies*, *22*(6), 2201–2238. https://doi.org/10.1093/rfs/hhn098.

Chordia, T., Subrahmanyam, A., & Anshuman, V. R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, *59*(1), 3–32. https://doi.org/10.1016/S0304-405X(00)00080-5.

Chung, K. H., & Zhang, H. (2014). A simple approximation of intraday spreads using daily data. *Journal of Financial Markets*, *17*, 94–120. https://doi.org/10.1016/j.finmar.2013.02.004.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. *Journal of Business & Economic Statistics*, *33*(1), 1–24. https://doi.org/10.1080/07350015.2014.983236.

Díaz, A., & Escribano, A. (2020). Measuring the multi-faceted dimension of liquidity in financial markets: A literature review. *Research in International Business and Finance*, *51*. https://doi.org/10.1016/j.ribaf.2019.101079.

Engle, R. F., Ghysels, E., & Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *The Review of Economics and Statistics*, *95*(3), 776–797.

Faff, R. W., Hillier, D., & Hillier, J. (2000). Time Varying Beta Risk: An Analysis of Alternative Modelling Techniques. *Journal of Business Finance & Accounting*, *27*(5–6), 523–554. https://doi.org/10.1111/1468-5957.00324.

Ferrara, L., & Marsilli, C. (2013). Financial variables as leading indicators of GDP growth: Evidence from a MIDAS approach during the great recession. *Applied Economics Letters*, *20*(3), 233–237. https://doi.org/10.1080/13504851.2012.689099.

Fong, K. Y. L., Holden, C. W., & Trzcinka, C. A. (2017). What are the best liquidity proxies for global research? *Review of Finance*, *21*(4), 1355–1401. https://doi.org/10.1093/rof/rfx003.

Foroni, C., Marcellino, M., & Schumacher, C. (2011). *U-MIDAS: MIDAS regressions with unrestricted lag polynomials* (Discussion Paper Series 1: Economic Studies, No 35). https://www.econstor.eu/bitstream/10419/55529/1/685618153.pdf.

Fuertes, A.-M., & Olmo, J. (2012). Exploiting Intraday and Overnight Price Variation for Daily VaR Prediction. *Frontiers in Finance and Economics*, *9*(2), 1–31.

Ghysels, E., Kvedaras, V., & Zemlys, V. (2016). Mixed frequency data sampling regression models: The R package midasr. *Journal of Statistical Software*, *72*(4), 1–35. https://doi.org/10.18637/jss.v072.i04.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). *The MIDAS Touch: Mixed Data Sampling Regression Models* (CIRANO Working Papers, 2004s-20). https://www.cirano.qc.ca/files/publications/2004s-20.pdf.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, *131*(1–2), 59–95. https://doi.org/10.1016/j.jeconom.2005.01.004.

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*(1), 53–90. https://doi.org/10.1080/07474930600972467.

Giot, P. (2005). Market risk models for intraday data. *European Journal of Finance*, *11*(4), 309–324. https://doi.org/10.1080/1351847032000143396.

Gold, N., Wang, Q., Cao, M., & Huang, H. (2017). Liquidity and volatility commonality in the Canadian stock market. *Mathematics-in-Industry Case Studies*, *8*(7), 1–20. https://doi.org/10.1186/s40929-017-0016-9.

Hautsch, N., & Jeleskovic, V. (2008). *Modelling High-Frequency Volatility and Liquidity Using Multiplicative Error Models* (SFB 649 Discussion Papers). http://dx.doi.org/10.2139/ssrn.1292493.

Hiemstra, C., & Jones, J. D. (1994). Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation. *Journal of Finance*, *49*(5), 1639–1664. https://doi.org/10.2307/2329266.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2019). *Forecast: Forecasting functions for time series and linear models. R package version 8.5*.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *27*(3), 1–22. https://doi.org/10.18637/jss.v027.i03.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Kim, J. H. (2014). *VAR.etp: VAR modelling: estimation, testing, and prediction. R package version 0.7*.

Kim, H. H., & Swanson, N. R. (2018). Methods for backcasting, nowcasting and forecasting using factor-MIDAS: With an application to Korean GDP. *Journal of Forecasting*, *37*(3), 281–302. https://doi.org/10.1002/for.2499.

Laurent, S., & Violante, F. (2012). Volatility forecasts evaluation and comparison. *WIREs Computational Statistics*, *4*(1), 1–12. https://doi.org/10.1002/wics.190.

Liang, S. X., & Wei, J. K. C. (2012). Liquidity risk and stock returns around the world. *Journal of Banking and Finance*, *36*(12), 3274–3288. https://doi.org/10.1016/j.jbankfin.2012.07.021.

Ma, R., Anderson, H. D., & Marshall, B. R. (2018). Stock market liquidity and trading activity: Is China different? *International Review of Financial Analysis*, *56*, 32–51. https://doi.org/10.1016/j.irfa.2017.12.010.

Ma, Y.-r., Ji, Q., & Pan, J. (2019). Oil financialization and volatility forecast: Evidence from multidimensional predictors. *Journal of Forecasting*, *38*(6), 564–581. https://doi.org/10.1002/for.2577.

Maas, B. (2019). Short-term forecasting of the US unemployment rate. *Journal of Forecasting*, *39*(3), 394–411. https://doi.org/10.1002/for.2630.

Monteforte, L., & Moretti, G. (2013). Real-time forecasts of inflation: The role of financial variables. *Journal of Forecasting*, *32*(1), 51–61. https://doi.org/10.1002/for.1250.

Ong, M. A. (2015). An information theoretic analysis of stock returns, volatility and trading volumes. *Applied Economics*, *47*(36), 3891–3906. https://doi.org/10.1080/00036846.2015.1019040.

Pástor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, *111*(3), 642–685. https://doi.org/10.1086/374184.

Pfaff, B. (2008a). *Analysis of Integrated and Cointegrated Time Series with R* (2nd edition). Springer.

Pfaff, B. (2008b). VAR, SVAR and SVEC models: Implementation within R package vars. *Journal of Statistical Software*, *27*(4), 1–32. https://doi.org/10.18637/jss.v027.i04.

Rösch, C. G., & Kaserer, C. (2014). Market liquidity in the financial crisis: The role of liquidity commonality and flight-to-quality. *Journal of Banking and Finance*, *37*(7), 2284–2302. https://doi.org/10.1016/j.jbankfin.2013.01.009.

Santos, D. G., & Ziegelmann, F. A. (2014). Volatility forecasting via MIDAS, HAR and their combination: An empirical comparative study for IBOVESPA. *Journal of Forecasting*, *33*(4), 284–299. https://doi.org/10.1002/for.2287.

Smith, P. (2016). Google's MIDAS touch: Predicting UK unemployment with Internet Search Data. *Journal of Forecasting*, *35*(3), 263–284. https://doi.org/10.1002/for.2391.

Tsui, A. K., Xu, C. Y., & Zhang, Z. (2018). Macroeconomic forecasting with mixed data sampling frequencies: Evidence from a small open economy. *Journal of Forecasting*, *37*(6), 666–675. https://doi.org/10.1002/for.2528.

Yeyati, E. L., Telia, D., & Schmukler, S. L. (2008). Emerging Market Liquidity and Crises. *Journal of the European Economic Association*, *6*(2–3), 668–682.