Forecasting Realized Volatility Using Machine Learning and Mixed-Frequency Data (The case of the Russian Stock Market)

Vladimir Pyrlik
Daniil Luchin

Higher School of Economics
Saint-Petersburg, November 2021

Forecasting Realized Volatility Using Machine Learning and Mixed-Frequency Data (the Case of the Russian Stock Market)

Vladimir Pyrlik Daniil Luchin

HSE University

November 2021

Abstract

We assess the performance of selected machine learning algorithms (lasso, random forest, gradient boosting, and long short-term memory) in forecasting the daily realized volatility of returns of selected top stocks in the Russian stock market in comparison with a heterogeneous autoregressive realized volatility benchmark in 2018-2020. We seek to improve the predictive power of the models by including various economic indicators that carry information about future volatility. We find that lasso delivers a good combination of easy implementation and forecast precision. The other algorithms require fine-tuning and frequent re-training, otherwise they are likely to fail to outperform the benchmark often enough. Only the basic lagged log-RV values are significant explanatory variables in terms of the benchmark in-sample quality. Many economic indicators of mixed frequencies improve the predictive power of lasso though, including calendar and overnight effects, financial spillovers from local and global markets, and various macroeconomics indicators.

Keywords: heterogeneous autoregressive model; machine learning; lasso; gradient boosting; random forest; long short-term memory; realized volatility; Russian stock market; mixed-frequency data

1 Introduction

Stock market volatility is known to be a measure of the dispersion of stock returns, and it is commonly used to assess the riskiness of an asset. For the majority of asset and investment risk management problems, volatility is one of the most important and irreplaceable characteristics of assets. Hence, forecasting stock returns volatility has been popular among financial market researchers.

Measuring volatility of returns is performed in various ways, with the *realized volatility* (RV) approach popular among practitioners and researchers. It has proven to be a preferred volatility evaluation technique due to its natural use of more information from high-frequent market data. Andersen and Bollerslev (1997) were the first to show that realized volatility forecasts outperform the predictions of numerous alternative approaches.

The concept of realized volatility was first introduced by Andersen and Bollerslev (1998) and is defined as the sum of squared intraday stock returns. An important advantage of the RV approach is that this volatility measure is observed directly from data, unlike the definitions of market volatility based on latent volatility models such as stochastic volatility (SV) or generalized autoregressive conditionally heteroskedastic models (GARCH).

Modeling and forecasting realized volatility can be done via several approaches, such as the heterogeneous autoregressive realized volatility (HAR-RV) model or multiplicative error model (MEM). Alternative approaches to modeling volatility are often compared. Andersen et al. (2003) show that HAR-RV is superior to SV models and GARCH. Hence, we use HAR-RV as the benchmark in our research.

In the literature, it is common to apply various modifications to the models to improve the quality of the forecasts. The method most commonly used and one that has proven relatively reliable is inclusion of exogenous variables that contain valuable information about the dynamics of volatility, the asset at hand or even the market or

the economy in general. Based on the literature, the most commonly used variables to explain the dynamics of volatility and to improve the predictive power of the models can be divided into the following groups.

Financial market variables are one of the most extensive and prominent sets of indicators in the literature. Lagged equity market returns are often shown to predict volatility. For example, a well known stylized fact on most markets is that, if market returns are negative, volatility increases (Christiansen et al., 2012; Nonejad, 2017). Earning-price ratio is an important indicator of a firm's wellbeing and value, so changes in the ratio potentially predict the stock returns volatility. When the earning-price ratio decreases, it is likely to indicate poor current and future performance of the firm, and hence a higher level of stock returns volatility in the future (Christiansen et al., 2012; Nonejad, 2017). Similarly, the dividend-price ratio can capture changes in stock returns volatility through the channel of investment productivity. When this ratio decreases, stock returns volatility is expected to increase (Christiansen et al., 2012; Nonejad, 2017). Long-term bond returns are considered to carry higher risks than short-term ones, and thus have higher interest rates. Variations of these quantities can be used to proxy investor attitude towards risk (Nonejad, 2017; Audrino et al., 2020).

Market liquidity is another important indicator that provides information about stocks returns and their volatility. An increase in liquidity is expected to indicate an increase in the level of market participants activity in the market. A significant change in activity typically leads to changes in price levels, returns, and volatility. As liquidity is not directly observable, a variety of indicators have been developed to capture it (such as Amihud, Roll, and High-Low). According to Będowska-Sójka and Kliber (2019), there is a significant relationship between volatility and liquidity, but the sign of the correlation can differ depending on the liquidity proxy. The authors conclude that the most relevant approximation of liquidity is High-Low, as this measure unilaterally influences volatility. Xu et al. (2019) show that there is a non-linear dependence

between liquidity and volatility with persistent influence of the former on the latter. This research also exhibits that High-Low liquidity proxies are the most influential in realized volatility forecasting.

Further, the number of daily transactions may carry significant information for movements of stock volatility. Some studies confirm this effect, while others state that it does not exist. For example, Shahzad et al. (2014) show that the number of trades in a day is a more significant predictor of volatility than average daily volume. Moreover, they demonstrate that the number of individual trades is a more important predictor than the number of trades by institutional market participants. A possible explanation is that individuals' actions represent a noise term (because they possess less reliable information about the market than organizations), which, in certain time periods, can lead to abnormally high volatility. Wang et al. (2015) also confirm the existence of the trading volume effect and point out that, the longer the forecasting horizon is, the lower the influence is. As for a contrary view of this effect, Todorova and Souček (2014) show that, for the German market, the trading volume of stock does not include any significant information for explaining realized volatility. It is worth noting that this result was achieved both in-sample and out-of-sample.

Stock volatility is also known to be significantly time-dependent. Hence, incorporation of day-of-the-week, weekend, and holiday effects is of great importance for precise forecasting of stock volatility. Many authors focus specifically on the effects of non-trading days. Martens et al. (2009) claim that stock volatility is usually higher after holidays, and on Christmas, half of its regular level. As shown by Wang and Hsiao (2010), weekday holidays increase the volatility of the S&P 100 and FTSE 100, while half-day trading periods decrease it. Diaz-Mendoza and Pardo (2020) find that volatility significantly decreases on the first day after a holiday or weekend, but after a long holiday, volatility either rises or remains the same.

Similarly, overnight and lunch-break periods are relevant for forecasting returns

volatility, because during these periods important information on trade or macroeconomic news may arrive. According to Wang et al. (2015), these non-trading periods significantly influence volatility. Moreover, they state that the leverage effect is captured, as the volatility rises higher after negative shocks to returns. The same results are achieved by Todorova and Souček (2014) and Zhu et al. (2017), who claim that the effects of overnight returns are higher than those of lunch-break returns.

In addition to the calendar effects, expiration-day effects of related derivatives have been thoroughly investigated. These effects measure how futures or option contract trading close to an expiration day may influence the underlying stock returns and volatility. This has been studied for stock markets in various countries, and the results are drastically different. Bollen and Whaley (1999) use Chinese data and do not discover statistically significant difference in stock volatility on expiration and non-expiration days of the derivatives. A similar result is achieved by Xu (2014) using Swedish data. However, Arago and Fernandez (2002) conclude that, for the Spanish market, volatility is significantly higher during a week with an expiration day. Chou et al. (2006) arrive at the same conclusion in the case of the Taiwanese market.

The inclusion of a variety of macroeconomic indicators is justified, because the overall economic environment influences the well-being of the corporate sector and thus the volatility in the market. The most frequently used proxies are CPI, industrial production growth, and GDP growth (Wongbangpo and Sharma, 2002; Christiansen et al., 2012; Paye, 2012; Nonejad, 2017; Audrino et al., 2020; Fang et al., 2020; Thampanya et al., 2020). It is important to note that, when macroeconomic variables are used in combination with financial indicators, most of the time, the former appear to be insignificant. Housing starts is one of the few indicators that has proven to influence volatility (Audrino et al., 2020; Fang et al., 2020). A possible mechanism behind this effect is that the more new houses are built, the more the credit market expands Fang et al. (2020). T-bill rates are often used as predictors of market volatility. If the econ-

omy is unstable, then T-bill rates generally tend to decrease, while volatility commonly increases. These variables are used to proxy the steadiness of the current economic situation (Christiansen et al., 2012; Nonejad, 2017; Audrino et al., 2020).

Not only does the domestic market affect stock volatility, but so do spillover effects from adjacent or global financial markets. These spillover effects represent the influence of foreign or adjacent markets on a local market. Balli et al. (2015) show that one of the important representations of the spillover effect is the trading volume of goods between developed and emerging markets. They also demonstrate that spillovers from the US are higher than those from Europe or Japan. Martens et al. (2009) illustrate that RV is higher on news announcement dates. Similarly, Wang and Hsiao (2010) demonstrate that for the Taiwanese market, weekend days raise the volatility of stocks, because, typically, a considerable amount of macroeconomic news is issued on Fridays in the US.

Further, the oil market appears to be closely connected to stock prices, which is a manifestation of spillover between adjacent markets. According to Kang et al. (2015) negative shocks in oil production lead to positive shocks in stock returns volatility. Similarly, an increase in demand for oil translates into a decrease in volatility. Luo and Qin (2017) state that oil price shocks positively influence returns on the Chinese stock market, as a rise in the oil price is a sign of an upturn in the economy.

Changing the functional form of the regression is another approach that is often used to improve both the explanatory power, and the predictive performance of the models. On the one hand, models like HAR-RV or MEM are commonly said to exhibit a relevant level of interpretability. On the other hand, they are not guaranteed to deliver reasonable forecasting power for either short or long time horizons, due to their limited ability to capture effects that are more complicated than the linear correlations between the volatility dynamics and the explanatory variables. However, machine learning (ML) algorithms are specifically known for highly accurate predictions, due to their ability to capture various non-linear patterns in the relationships between the variables. Recently,

much attention has been focused on investigating the applicability of ML in forecasting returns and their volatility.

Ingle and Deshmukh (2021) implement several types of models to predict closing prices of stocks: Generalized Linear Model (GLM), Gradient Boosting Model (GBM), and several types of neural networks in combination with machine learning methods. The results show that GLM displays the highest level of forecasting accuracy, followed by ensemble models and deep learning networks.

Hamid and Iqbal (2004) use a three-layered neural network to forecast the volatility of S&P 500 futures and show that the predictions significantly outperform the benchmark. Further, Parisi et al. (2008) research changes in the market price of gold and find that the best performance, in-sample and out-of-sample, is delivered by a rolling neural network. Ding et al. (2015) investigate potential improvements in predictions for S&P 500, and show that forecasts from a deep convolutional neural network appear 6% better than those from the baseline model.

Long Short-Term Memory (LSTM) is another machine learning method that has been gaining popularity among researchers and practitioners. Xiong et al. (2015) compare the S&P 500 returns volatility forecasts from GARCH, Lasso regression, Ridge regression, and LSTM. The results show that the LSTM forecasts significantly outperform its competitors. Another notable study is Liu (2019), which shows that a combination of recurrent neural networks with LSTM significantly outperforms GARCH in forecasting the returns volatility of S&P 500 and AAPL.

Combining multiple models into one appears to be an effective and, hence, popular approach to applying deep learning algorithms. For example, Kristjanpoller and Minutolo (2015) use artificial neural networks to combine GARCH-based forecasts of the gold price returns variance, and achieve a sound reduction in the mean average percentage error. Vidal and Kristjanpoller (2020) combine LSTM and convolutional neural networks for gold price returns volatility forecasting, achieving a significantly

better predictions than GARCH or LSTM alone.

An important feature of the current literature on volatility forecasting, using either traditional approaches or ML, is the scope of the markets under analysis. Overall, most research focuses on the US, Chinese and European markets. Few studies consider emerging markets, and even fewer consider Russia (Aganin et al., 2017; Nagapetyan et al., 2019; Fantazzini and Shangina, 2019; Fantazzini, Fantazzini; Bazhenov and Fantazzini, 2019; Aganin, 2020). To the best of our knowledge, no research on the Russian stock market has gone beyond GARCH-type or HAR-RV-type of methodology in the analysis of returns volatility. Hence, our main goal is to contribute to the existing literature by performing a comparative analysis of several approaches to forecasting RV in the context of the Russian stock market, including the HAR-RV and ML approaches.

We first aim to identify the extent to which ML is more suitable for RV forecasting than the benchmark HAR-RV on the Russian stock market. Secondly, we seek to learn, what information is significant for the Russian stock market RV forecasting, and how it is different from the situation on international markets. To achieve these goals, we extract an extensive dataset for the Russian stock market and compare the out-of-sample performance of the HAR-RV and 4 ML algorithms (Lasso, Random Forest, Gradient Boosting, and Long Short-Term Memory) in returns RV forecasting for selected top stocks in the market. We find that both the HAR-RV and ML approaches provide us with reasonable predictive power in terms of RMSE of RV in a rolling forecasting scheme, with the ML generally outperforming the benchmark when a reasonable set of exogenous features are included. In particular, Lasso regression appears to deliver a convenient combination of easy implementation and forecasts precision. More complicated algorithms (Random Forest, Gradient Boosting, Long Short-Term Memory) are very promising, but we show that, to benefit from them, they require fine-tuning and frequent re-training, which is a computationally demanding task.

The rest of the paper is organized as follows. Section 2 introduces the methodology

of the benchmark HAR-RV model and ML algorithms used in this research, and the data splitting and forecasting schemes we choose. In Section 3, we describe the dataset and proceed to our exploratory data analysis in Section 3.2. In Section 4, we describe the modeling technique and analyze the results in Section 5. Section 6 offers a discussion of our main findings and limitations of the research, and Section 7 concludes. Supplementary aids on the data and results are collected in an online appendix.

2 Methodology

2.1 The Benchmark Model

The benchmark model in our study is HAR-RV of Corsi (2009). The main idea of the approach is to use high-frequency data to obtain more accurate forecasts of volatility based on daily, weekly, and monthly RV. The notation of the model is:

$$RV_{t+1d}^{(d)} = \alpha + \beta^{(d)} \cdot RV_t^{(d)} + \beta^{(w)} \cdot RV_t^{(w)} + \beta^{(m)} \cdot RV_t^{(m)} + \omega_{t+1d}, \tag{2.1}$$

where weekly realized volatility, for example, is given by

$$RV_t^{(w)} = \frac{1}{5} \cdot (RV_t^{(d)} + RV_{t-1d}^{(d)} + ... + RV_{t-4d}^{(d)}). \tag{2.2}$$

Inclusion of additional regressors to the model is straightforward:

$$RV_{t+1d}^{(d)} = \alpha + \beta^{(d)} \cdot RV_t^{(d)} + \beta^{(w)} \cdot RV_t^{(w)} + \beta^{(m)} \cdot RV_t^{(m)} + \sum_i \beta_i \cdot x_{it} + \omega_{t+1d}, \tag{2.3}$$

where x_{it} is an additional explanatory variable i at moment t.

Estimation of the model is typically performed via OLS. Newey-West robust standard errors are used to retain consistency of estimates with heteroskedicity and autocorrelation of the error term. When the extended version (2.3) of the model is used,

a model selection technique is required for an in-sample based selection of the optimal combination of explanatory variables. We estimate several specifications of the model with differing additional explanatory variables and select the one that minimizes AIC from those without significant residual autocorrelation.

2.2 Machine Learning Algorithms

Here, we briefly describe the ML algorithms we use and compare to the benchmark HAR-RV. These are: Lasso regression (Lasso), Gradient Boosting (GB), Random Forest (RF), and Long Short-Term Memory (LSTM). Below, we cover distinct features of each algorithm, with a fuller description and explanations of the mechanisms given in Appendix A1, page 33.

The main feature of Lasso is regularization on weights of linear regression with zeroing of extreme-value coefficients. This algorithm is typically good at dealing with overfitting that may occur as a result of either a relatively small sample size or too many collinear regressors. Lasso uses the only hyperparameter, which is the penalty for the degree of collinearity. Hence, training this algorithm is fast.

GB is an ensemble algorithm with a consequent learning of regression trees, while RF is another ensemble algorithm that uses parallel learning of regression trees. Due to the consequent structure, GB is capable of accurate capturing of dependencies in the data, but is prone to overfitting. RF, on the other hand, is more robust to overfitting. However, both algorithms are considered well suited for feature selection and coping with multicollinearity. When some features appear highly collinear, the trees will avoid using them together for the sake of greater information gain. These algorithms classify some variables as the most/least significant, depending on their inputs to information gain.

As for neural networks, LSTM is a type of RNN that works with sequences of variables. Due to its recurrent structure, the algorithm can capture autoregressive

dependencies, which makes it particularly useful in the tasks of time series forecasting. In contrast with other networks, LSTM is designed to work better with longer sequences of data. The architecture of LSTM is tunable, which makes the algorithm flexible for different data types and tasks. The quality of this algorithm also depends on the learning approach. Hence, such hyperparameters as batch size, learning rate, number of epochs, and type of the optimizer should also be tuned. LSTM, thus, is the most complicated and computationally challenging algorithm among those used in our study.

2.3 Partitioning the Data and Training the Models

To train and evaluate our models, we perform a rolling scheme with out-of-sample validation and testing. We sequentially divide our dataset into training, validation, and test sub-samples. Each training sample includes information over a two-year period, and the validation and test samples are the following two quarters (one each). We roll forward by one quarter at a time. We use RMSE to measure the quality of our forecasts.

The validation parts are used to select the hyperparameters of the models (when there are any). The choice of the hyperparameters is made via grid search over excessive sets of possible values of the parameters. The optimal combinations are chosen to minimize the validation RMSE.

3 Data

3.1 Groups of Variables

We extract historical data on the stock prices of the top 9 companies of Moscow Exchange index from the 1st of January 2016 to the 31st of December 2020 at 5-minute frequency[1]. We use the data to calculate daily, weekly and monthly stock realized

[1]The data is open and can be obtained directly from the stock exchange website or another service; we obtain the data using the stock prices historical data exports feature of FINAM, www.finam.ru

volatility for each company.

Following the results of previous studies, we add a variety of explanatory variables to our dataset.

- To give an alternative for the T-bill rate, daily values of the Russian Government Bond Index (RGBI) are included. Because returns on market portfolios cannot be reported directly, daily log-returns on the stock market index RTSI were taken as a proxy. An important characteristic for this proxy is high level of diversification; RTSI is a composite index with the most liquid Russian stocks[2].

- To control for changes in the economic environment and macroeconomic circumstances, we added the dynamics of GDP (quarterly), CPI (monthly), and dwellings commenced (monthly)[3]. From the same source, we obtained a few financial performance indicators, specifically, the dynamics of dividend price ratio and earning price ratio for each of the 9 companies (monthly). It is important to note that for POLYUS (www.polyus.com), the major part of these variables do not appear to be available; hence, we omit them from the specifications for POLYUS.

- We included exports and imports to/from the USA from/to Russia, using the data from the census.gov WebSite. In the literature, the exports and imports to/from the USA are classified as the spillover effect. However, in our research, we include them into the group of macroeconomic indicators. This is due to the frequency of these variables, and the fact that the mechanism of the spillovers is typically explained in terms of macroeconomic theory, for example, Balli et al. (2015).

- We calculate several market liquidity indicators: High-Low, Amihud, and Roll, following the approach of Będowska-Sójka and Kliber (2019). However, we end

[2]The data on RGBI and RTSI are available at www.finam.ru

[3]These data can be obtained from the Refinitiv Eikon (Thompson Reuters) https://eikon.thomsonreuters.com

up including only the High-Low metrics into our models, since it showed the most effect on stock volatility in the literature.

- We accounte for the holiday effect, the weekend effect, and the Friday effect by including the respective dummy variables. To try to capture an eponymous effect, we add the overnight returns to the sets of variables, following the approach of Wang et al. (2015).

- Finally, we included the realized volatility of S&P 500 and Brent oil price to reflect spillover effects from the global stock and crude oil markets.

It is worth noting that the companies in our study represent various sectors of the economy: banking, mining, retail, and oil and gas. Literature shows that spillovers between sectors are of great importance. As presented by Hammoudeh et al. (2009), three main sectors (industrial, service, and banking) of GCC economies demonstrate volatility spillovers. Chen et al. (2019) confirmed results of the previous paper and showed that consumer discretionary, industrial, and health sectors generate the largest spillovers. The US stock market also features cross-volatility between sectors, as shown by Mensi et al. (2020). They demonstrated that consumer services and goods sectors produces the largest amount of volatility, while material sectors produce the least.

Due to the industrial specificity of the Russian economy, it happens that most companies chosen for our research belong to the oil and gas sector. Hence, we do not expect to see much evidence of volatility spillovers between sectors. However, this is a field for future research.

We now describe the specificity of the data, necessary transformations of the variables, and creation of additional indicators, when required. To avoid negative forecasts of realized volatility, we apply the natural logarithm to the dependent variable and its lagged values. We shift to the growth rates of the low-frequent variables to introduce more variation in our data and achieve stationarity. We include the lagged series of the

main variables into our datasets. As a result, we divide all our variables into 5 groups (see Table 1) to investigate additional predictive power that each group of variables brings to a certain model.

Table 1: Groups of explanatory variables included into models

Group	**Variables**
Basic	log RV, log RV weekly, log RV monthly
Overnight and calendar effects	is after weekend, is after holiday, is Friday, overnight returns
Financial effects	growth rate of dividend price ratio†, growth rate of earning price ratio†, High-Low, log-returns of RTSI, RGBI
Spillovers	log RV of S&P, log RV of Brent
Macro indicators	growth rate of import/exports from/to the USA†, growth rate of CPI†, growth rate of housing starts†, growth rate of GDP†

† - low frequency variable

We then construct 5 specifications of all implemented models with the consecutive addition of these groups of variables and 5 specifications with lags of variables. We have some missing values in the data. To keep the datasets as complete as possible, for each particular company we omit variables which are missing at rates 30% or more in the training samples. Less frequent missing values are replaced with the latest known values of the same variable.

3.2 Exploratory Data Analysis

We conduct a preliminary data analysis to identify general patterns within the data, to detect possible effects of extreme events, and to evaluate the overall relationships between the variables.

Firstly, we consider the dynamics of the logarithm of realized volatility to investigate changes throughout the period; see Figure 1 for an example (the rest of the figures are in Appendix A2.1, page 37). Overall, our dependent variable is a typical time

series of this kind: a volatile and possibly heteroskedastic series, yet most likely with a stable longer run average level and range. Visually, there are specific differences across companies, and some common patterns. For example, all the series show significant short-run increases in volatility in the first half of 2018, and in the first half of 2020, there is an obvious and very sound change in the average volatility level. The shifts in 2018 are likely to be the result of GDP growth deceleration due to sanctions policies of foreign countries, and depreciation of the national currency. The shifts in 2020 are obvious consequences of the COVID-19 crisis, and of the oil market shocks. We address sampling around those periods with caution, yet we expect loss in the predictive power of the models anyway.

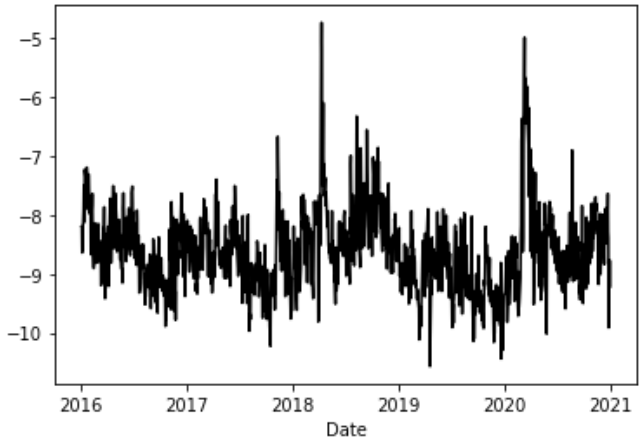

Figure 1: Dynamics of logarithm of realized volatility of returns, SBERBANK

Secondly, as we aim to capture changes in mean of our dependent variables that occur over time, we consider the distribution of the average level of the dependent variable across weekdays (see Appendix A2.2, page 40). For all companies, Thursdays feature the highest average volatility level. We also observe that, for all companies, except GAZPROM and POLYUS, RV is the lowest on Mondays on average. For GAZPROM and POLYUS, Friday features the lowest RV. These findings line up with the literature. We include the corresponding dummies into our datasets.

Next, we select variables that are company-specific according to the information that we want to account for in our models. We present and analyze the descriptive statistics, including sample averages, standard deviations, and autocorrelations (see Appendix A2.3, page 43). For all companies, RV and High-Low proxy of liquidity exhibit persistent significant autocorrelation. Growth rates of dividend price ratio and earning price ratio show only a few significant autocorrelation terms, and this order appears to be company-specific.

Finally, to examine correlations between the dependent variable and some potential explanatory variables, we examine correlation-scatterplot diagrams (see Appendix A2.4, page 46). For each company, we select the following indicators: logarithm of stock realized volatility (the dependent variable), logarithm of realized volatility of S&P 500 (captures potential spillovers from the global market), High-Low liquidity proxy (liquidity spillovers), RGBI (proxy of the economy steadiness), and the log-returns of RTSI (the local market spillovers). The results are rather similar across the companies. The log-RV is significantly correlated with the other variables; it approves the inclusion of these indicators in the models. However, the there are obvious non-linear dependencies present, hence, machine learning algorithms are naturally expected to perform well.

Overall, the RV on the Russian stock market features properties typically outlined in the literature. Also, this measure appears to be relatively similar across the companies with drastic changes of values early in 2018 and 2020. We also find weekday effects in the average level of RV. Further, we observe sound correlation between RV and various indicators that we intend to use in the models as explanatory features. However, these dependencies are ambiguous. We need to conduct a thorough search for the optimal forecasting specification in the case of each particular asset and over a particular period of time.

4 Modeling Technique

We aim to forecast realized volatility on the Russian stock market as precisely as possible, and to identify the effects of potentially informative variables on the volatility dynamics. We perform our analysis iteratively, going over different predictive algorithms combined with different groups of explanatory variables. Each such combination is estimated for each of the assets in our datasets on rolling samples. The performance of the different models is then compared across time, across sets of explanatory variables, across algorithms, and across assets.

For each dataset we perform the following sequence of steps.

1. We split the data across time into 11 samples, using the rolling window scheme, rolling forward by one quarter at a time. Each sub-sample consists of 10 consecutive quarters.
   (a) The first two years (8 quarters) are used to train the models.
   (b) The next one quarter is a validation sample, used to tune the hyperparameters of the models.
   (c) The succeeding one quarter is the testing sample, used to assess the models' performance via comparing the RMSE of the forecasts of log-RV.
2. We construct 40 different model specifications, combining different model types with sets of explanatory variables.
   (a) HAR-RV is our benchmark model, and our ML algorithms are: Lasso, RF, GB, and LSTM.
   (b) We split all the explanatory variables into 5 groups (see Table 1 in Section 3), and for every model specification we add groups consecutively.
   (c) Each of the algorithms is trained on the sets of variables either including lagged values or not, except in the case of HAR-RV and LSTM.

3. All the models are trained in terms of minimization of RMSE. The hyperparameters of the models are tuned via a grid search to minimize RMSE, too, but on the validation sample.

4. We choose RMSE as the main measure of the predictive power, too.

5. In addition to plain comparison of the RMSE of different models, we consider relative win-rates of different specifications by algorithm and by set of variables. We track the velocity of the accumulation of the squared sum of forecast errors on the testing samples, to gather insights about which models and which sets of variables should be preferred, and how they can be further improved.

5 Results

5.1 HAR-RV and the Explanatory Power of the Variables

The full results of the estimation of the benchmark HAR-RV regressions are available in Appendix A3, page 55. Overall, the most commonly selected statistically significant variables are the lagged daily, weekly, and monthly log-RV, the liquidity proxy High-Low, and the log-RV of S&P 500. Regarding the other variables, growth rate of exports and log-returns of RTSI appear to be significant for some firms (3 of 9). The variables that are not significant in any of the regressions are from the group of calendar effects. The signs of the significant coefficients coincide with those found in other studies which considered similar effects.

Regarding the global market effects, the result is that, the higher the S&P 500 RV is, the higher the stock realized volatility of a Russian company tends to be. This confirms the existence of spillover effect from the global market. The liquidity proxy High-Low also has a reasonable sign: the higher the liquidity of a stock is, the lower is its volatility. As for the basic variables, their signs make sense as well, because the

higher the weekly or monthly volatility is, the higher the value of the dependent variable is. The log-returns on RTSI (the local market effect) show a negative effect, similarly to the findings of Christiansen et al. (2012).

Though the calendar effects are not significant, the signs of their coefficients also coincide with those typically found in the literature. Similarly, for example, to the results of Diaz-Mendoza and Pardo (2020) and Todorova and Souček (2014), we find that, in the Russian market, too, volatility decreases after a holiday or a weekend, and due to high overnight returns.

We run Breusch-Godfrey LM tests for residual autocorrelation, and find that the results vary across the firms. There are companies for which there is significant residual autocorrelation in all the specifications (ROSNEFT, NORNICKEL, POLYUS, and MAGNIT). For the other firms, the residual correlation vanishes with inclusion of rather few additional variables (GAZPROM, LUKOIL and NOVATEK). Finally, for SBERBANK and POLYMETAL there are no signs of residual autocorrelation in all specifications of HAR-RV.

We compare the AICa of the specifications, and find that for most companies the specification of HAR-RV that includes spillover effects appears to be the best. This pattern is violated only for LUKOIL, for which regression with inclusion of all variables should be chosen. In general, this result has proven that the selected explanatory variables contain valuable information for explanation of stock realized volatility.

5.2 ML and Predictive Power of the Models

To report results of machine learning algorithms, we select top-1 models of each type in terms of average RMSE, and put them on one graph for each company (see Appendix A4.1, page 64). The most distinct features for all figures are peaks in RMSE early in 2020 and in different quarters of 2018 and 2019. The most reliable explanation, in our opinion, is the market shock from COVID-19 early in 2020, and the oil market shock in

the same time. For the 2018-2019 shocks, there could be multiple reasons, most likely including deceleration in growth of GDP due to sanctions policies of foreign countries, pension reform, and depreciation of the national currency in 2018 and 2019. However, in the quarter after those peaks, the RMSE lowers significantly, which indicates that the proposed models adjust to the new information, process it, and can regain their predictive power.

Further, for most of the companies, GB and RF algorithms appear to be the weakest. GB appears to be the worst overall in most cases (across time, across specifications, and across assets). This suggests that consequent learning of regression trees might not be the best for forecasting stock realized volatility. However, unlike the others, RF models have the lowest RMSE for all test periods for POLYUS. In turn, Lasso and HAR-RV appear to be the best models, replacing each other in the leading position in different test quarters. The benchmark model is chosen in its basic specification most of the time, while Lasso performs better with inclusion of all types of variables into the model. The model specifications without lags of variables are chosen more often, meaning that lags do not lend much forecasting power into the algorithms.

To understand the predictive capabilities of different models better, we considered top-3 specifications for each class of ML algorithms for each dataset (and the top-1 benchmark model specification for comparison). The predictive performance measures and description of the specifications are in Appendix A4.2, page 69. As in the previous step, Lasso and HAR-RV deliver the lowest average RMSE on the testing samples, followed by RF and GB. For LSTM, most notably, with the inclusion of extra variables, an increase in RMSE is much higher than for the other models. We believe that the poor performance of LSTM is a sign of overfitting. Nevertheless, it should be noticed that any model can deliver the lowest RMSE in a particular quarter. Thus, it is important that the majority of top-3 specifications of each model are based on variables without lags. Moreover, we conclude that even though top-1 specifications can be based on the basic

variables only, among the top-3 specifications there is commonly at least one model with addition of extra variables, which does not perform significantly differently across the top-3 options. This proves that various groups of variables indeed carry valuable information about volatility and are important for better forecasting. To confirm this claim further, we repeat the analysis of the top-3 specifications, excluding the influence of the 1st and 2nd quarters of 2020. The results are comparable to those from the previous step.

Since any algorithm can perform best in particular test quarters, we continue analysis of the results and compare the win-rate of the models by class, showing the number of cases among the firms for which a particular class of models appears to be the best in each particular quarter and on average overall. See Table 2. The most frequent winner is Lasso, followed by HAR-RV and RF. Hence, this result overlaps previous findings for benchmark model and Lasso. However, the result for RF demonstrates that, though RF does not appear as the top-1 model, it is still a powerful algorithm. LSTM and GB have the lowest win-rates. However, the two periods when either LSTM or RF show the highest win-rate are the 3d quarter of 2018, and the 3d quarter of 2020, right after the periods with abnormally high volatility for most of the companies. This suggests that these particular algorithms can adjust faster to changes in the patterns and absorb new arriving information better than the other models. If so, it is worth an attempt to improve their performance by more frequent re-training (more on this in the conclusion).

5.3 Prediction-Based Importance of the Variables

Because Lasso is among the best model classes in terms of prediction across both time periods and assets, we are able to determine which variables are the most significant. We point out two groups of relatively significant variables: those that were sustainably chosen by the algorithm, and those that were impermanently, but frequently chosen.

Table 2: Total win-rate of models by class and testing sample period

Period	**Models**				
	HAR-RV	*Lasso*	*LSTM*	*RF*	*GB*
2018Q2	2	**4**	0	1	3
2018Q3	1	2	1	**3**	2
2018Q4	2	**6**	0	1	0
2019Q1	**3**	2	1	2	1
2019Q2	1	2	2	2	2
2019Q3	**4**	3	0	2	0
2019Q4	2	**6**	0	1	0
2020Q1	2	**5**	0	2	0
2020Q2	**4**	**4**	0	1	0
2020Q3	2	1	**3**	**3**	0
2020Q4	**3**	**3**	0	2	1
Average	2.36	**3.45**	0.64	1.82	0.82

The groups are presented in tables in Appendix A4.3, page 73, and Table 3 below summarizes the results across all firms.

Table 3: Best overall variables, chosen by Lasso

	Group of variables
Sustainably chosen	Log RV, log RV weekly, log RV monthly, is after weekend, is Friday
Frequently chosen	Log-RV of S&P, log-RV of Brent, growth rates of imports, growth rates of exports, growth rates of GDP, growth rates of CPI, overnight returns, RGBI, earning price ratios, dividend price ratios, growth rates of housing starts, High-Low

According to these results, the first group includes basic HAR-RV model variables: logarithms of daily, weekly, and monthly realized volatility. It happens because the HAR-RV gives an accurate description of the autoregressive process of RV, and with so few variables the Lasso must be very close to the baseline linear model. Note that, even though the calendar effects are often insignificant in-sample, they are rather persistently chosen to improve prediction by Lasso (e.g., the Friday effect).

The second group contains less frequently chosen variables, including indicators of spillover effects (log-RV of S&P 500 and Brent oil price returns). In many cases,

macroeconomic factors including growth rates of GDP, CPI, imports and exports, and housing starts are significant. Moreover, financial indicators including growth rates of earning price ratios, growth rates of dividend price ratios, growth rates of housing starts, High-Low proxy of liquidity, and RGBI are important in forecasting realized volatility for most companies. Lastly, overnight returns were frequently chosen by Lasso. Similarly to the calendar effects, many of these variables are not detected as carriers of significant explanatory power in-sample by the benchmark.

Compared to the results obtained by the benchmark model, many more variables from various groups are chosen by Lasso. However, logs of daily, weekly, and monthly RV and High-Low are chosen by both algorithms.

Our results demonstrate that linear models are more suitable for RV forecasting than more complicated machine learning algorithms, at least in our framework. However, Lasso provides more accurate forecasts than HAR-RV. Importantly, in terms of predictive power optimization, Lasso tends to choose more variables as being valuable, while the benchmark model works the best on the basic sets of regressors.

There are multiple reasons for the relative failure of GB, RF, and LSTM in the task of RV forecasting. We believe that the most crucial source of high prediction numbers of errors by these algorithms is overfitting. Another possible reason for the failure of these algorithms is re-training that is not frequent enough. In both cases, the underlying reason must be the nature of the volatility process itself, as it is essentially noisy. It is less likely but possible that the tree structure used by GB and RF may be unsuitable for forecasting time series such as RV. Lastly, LSTM is the closest to HAR-RV and Lasso in terms of predictive power, but the problem of overfitting is likely to have escalated for this model.

6 Discussion

6.1 Applications of the Results

There are several ways the results of this study can be implemented. Firstly, we were able to identify the most suitable model for forecasting realized volatility on the Russian stock market, so researchers and investors who want to study this topic or trade on the market can use the model. Secondly, if researchers or investors want to build other models for forecasting stock volatility, they can use our findings on the significant predictors of realized volatility. Thirdly, with help of our results, traders can quantitatively assess the future short-run risks of an asset on the Russian stock market. Lastly, as realized volatility is important for optimal portfolio allocation, our results can be used by portfolio investors to improve their (re)allocation decisions.

6.2 Limitations of the Study

We encountered the impossibility of acquiring some data. Although we included variables related to calendar effects, spillover effects, and financial and macroeconomic effects, many factors that can influence RV could not be taken into account. The most obvious reason is unavailability of data. For instance, investors sentiment is expected to be an important predictor of RV, yet we have not yet managed to access suitable structured or unstructured data that would have been convenient to use in our research.

Another challenge is the computational capacity requirements necessary for appropriately fine and frequent tuning of the models, particularly, the computationally heavy ML algorithms (RF, LSTM). We briefly studied the velocity with which the sum of the squared prediction error is accumulated by different models within a particular testing quarter. Figure 2 shows the accumulation process for SBERBANK volatility top-1 by-class predictors during the 1st quarter of 2019. The overall winning model for this firm and this period was Lasso (it shows the lowest accumulated sum at the end of the

quarter, day 60, see the left panel of Figure 2). However, the superiority of this model is not stable within the period. Obviously, LSTM and GB have higher values of the squared error to begin with. This supports our intuition about the overfitting problem. However, the other three models start off rather close to each other in the beginning of the quarter, with RF being a sound leader for several days (see the right panel of Figure 2). RF becomes outdated rather quickly (after approximately 6 days), and does not recover throughout the rest of the testing sample. Moreover, the accumulation of the sum of the squared error occurs, on average, with increasing rates for all the models, with dramatic increases in some periods, which possibly signal arrivals of new information not yet accounted for by the trained models. These observations support the idea that more frequent re-training of the models might significantly improve predictive power. Even though this seems a rather obvious path to take, we leave it for future research, as it is too computationally demanding, particularly in the case of the GB, RF, and LSTM algorithms.

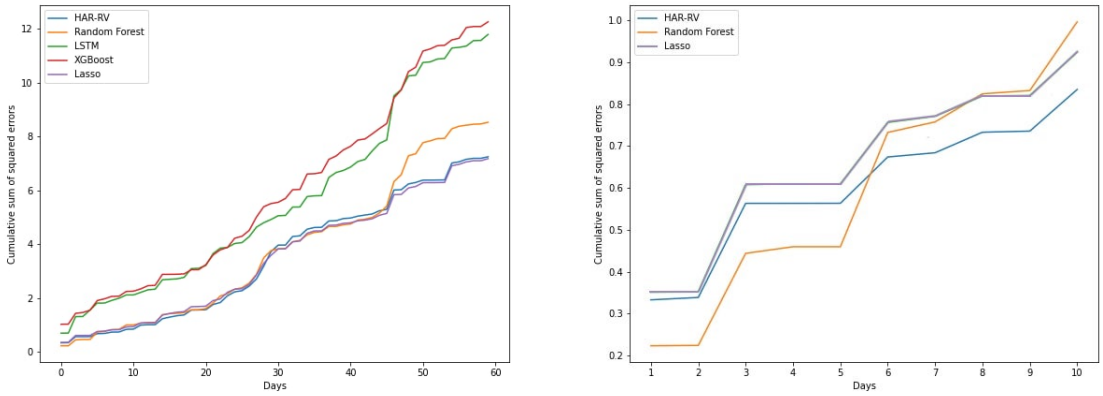

Figure 2: Top-1 predictors cumulative sum of squared errors, SBERBANK, 2019Q1

7 Conclusion

We aim to employ data on the Russian stock market and to compare the suitability of the benchmark HAR-RV with several ML algorithms (Lasso, Random Forest, Gradient Boosting, Long Short-Term Memory) in the task of forecasting daily RV of selected top stocks on the Russian stock market. We further seek to identify the most valuable factors for explaining the dynamics and forecasting the future values of the RV.

We collect a novel and extensive dataset for the top-9 Russian companies based on MICEX, consisting of our variable of interest and various groups of additional variables, including calendar effects, financial variables, spillover effects and macroeconomic variables. For each of the models, we constructed a number of specifications based on either HAR-RV or ML algorithms that are trained on various sets of explanatory variables.

The results show that Gradient Boosting, Random Forest, and LSTM did not appear to perform well in the forecasting task. The best performing models were Lasso and HAR-RV. From Lasso, we were able to highlight the most significant factors for forecasting the RV. The variables that showed the most effect on future RV across all companies are: logarithm of daily, weekly, and monthly realized volatility, High-Low proxy of liquidity, calendar effects, and some macroeconomic variables and financial market and spillover effects, including as, for example, logarithm of realized volatility of S&P, growth rates of CPI and GDP, growth rates of earning price ratio and dividend price ratio.

We also find that, once trained, the specifications become outdated rather quickly. Their predictive performance could be improved by finer tuning and more frequent re-training. This is a computationally heavy task, which could be addressed in future research. Furthermore, as most companies in our study are from the oil and gas sector (because of the industry specificity of the Russian economy), spillovers between sectors could not be investigated fully, opening another promising direction for further research.

References

Aganin, A. et al. (2017). Forecast comparison of volatility models on russian stock market. *Applied Econometrics 48*, 63–84.

Aganin, A. D. (2020). Russian stock index volatility: Oil and sanctions. *Voprosy Ekonomiki (2)*, 86–100.

Andersen, T. G. and T. Bollerslev (1997). Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *The journal of Finance 52 (3)*(3), 975–1005.

Andersen, T. G. and T. Bollerslev (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, 885–905.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71 (2)*(2), 579–625.

Arago, V. and A. Fernandez (2002). Expiration and maturity effect: empirical evidence from the spanish spot and futures stock index. *Applied Economics 34 (13)*(13), 1617–1626.

Audrino, F., F. Sigrist, and D. Ballinari (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting 36 (2)*(2), 334–357.

Balli, F., H. R. Hajhoj, S. A. Basher, and H. B. Ghassan (2015). An analysis of returns and volatility spillovers and their determinants in emerging asian and middle eastern countries. *International Review of Economics & Finance 39*, 311–325.

Bazhenov, T. and D. Fantazzini (2019). Forecasting realized volatility of russian stocks using google trends and implied volatility. *Russian Journal of Industrial Economics 12 (1)*(1), 79–88.

Będowska-Sójka, B. and A. Kliber (2019). The causality between liquidity and volatility in the polish stock market. *Finance Research Letters 30*, 110–115.

Bollen, N. P. and R. E. Whaley (1999). Do expirations of hang seng index derivatives affect stock market volatility? *Pacific-Basin Finance Journal 7 (5)*(5), 453–470.

Breiman, L. (2001). Random forests. *Machine learning 45 (1)*(1), 5–32.

Chen, Y., W. Li, and F. Qu (2019). Dynamic asymmetric spillovers and volatility interdependence on china's stock market. *Physica A: Statistical Mechanics and its Applications 523*, 825–838.

Chou, H. C., W. N. Chen, and D. H. Chen (2006). The expiration effects of stock-index derivatives: Empirical evidence from the taiwan futures exchange. *Emerging Markets Finance and Trade 42 (5)*(5), 81–102.

Christiansen, C., M. Schmeling, and A. Schrimpf (2012). A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics 27 (6)*(6), 956–977.

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

Diaz-Mendoza, A.-C. and A. Pardo (2020). Holidays, weekends and range-based volatility. *The North American Journal of Economics and Finance 52*, 101–124.

Ding, X., Y. Zhang, T. Liu, and J. Duan (2015). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Fang, T., T.-H. Lee, and Z. Su (2020). Predicting the long-term stock market volatility: A garch-midas model with variable selection. *Journal of Empirical Finance 58*, 36–49.

Fantazzini, D. Forecasting and backtesting of market risks in emerging markets. *Risk Assessment and Financial Regulation in Emerging Markets' Banking: Trends and Prospects*, 199.

Fantazzini, D. and T. Shangina (2019). The importance of being informed: forecasting market risk measures for the russian rts index future using online data and implied volatility over two decades. *Applied Econometrics 3 (55)*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Hamid, S. A. and Z. Iqbal (2004). Using neural networks for forecasting volatility of s&p 500 index futures prices. *Journal of Business Research 57 (10)*(10), 1116–1125.

Hammoudeh, S. M., Y. Yuan, and M. McAleer (2009). Shock and volatility spillovers among equity sectors of the gulf arab stock markets. *The Quarterly Review of Economics and Finance 49 (3)*(3), 829–842.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9 (8)*(8), 1735–1780.

Ingle, V. and S. Deshmukh (2021). Ensemble deep learning framework for stock market data prediction (edlf-dp). *Global Transitions Proceedings 2 (1)*, 47–66.

Kang, W., R. A. Ratti, and K. H. Yoon (2015). The impact of oil price shocks on the stock market return and volatility relationship. *Journal of International Financial Markets, Institutions and Money 34*, 41–54.

Kristjanpoller, W. and M. C. Minutolo (2015). Gold price volatility: A forecasting approach using the artificial neural network–garch model. *Expert systems with applications 42 (20)*(20), 7245–7251.

Liu, Y. (2019). Novel volatility forecasting using deep learning–long short term memory recurrent neural networks. *Expert Systems with Applications 132*, 99–109.

Luo, X. and S. Qin (2017). Oil price uncertainty and chinese stock returns: New evidence from the oil volatility index. *Finance Research Letters 20*, 29–34.

Martens, M., D. Van Dijk, and M. De Pooter (2009). Forecasting s&p 500 volatility: Long memory, level shifts, leverage effects, day-of-the-week seasonality, and macroeconomic announcements. *International Journal of Forecasting 25 (2)*(2), 282–303.

Mensi, W., R. Nekhili, X. V. Vo, T. Suleman, and S. H. Kang (2020). Asymmetric volatility connectedness among us stock sectors. *The North American Journal of Economics and Finance 56*, 101327.

Nagapetyan, A. et al. (2019). Precondition stock and stock indices volatility modeling based on market diversification potential: Evidence from russian market. *Applied Econometrics 4 (56)*, 45–61.

Nonejad, N. (2017). Forecasting aggregate stock market volatility using financial and macroeconomic predictors: Which models forecast best, when and why? *Journal of Empirical Finance 42*, 131–154.

Parisi, A., F. Parisi, and D. Díaz (2008). Forecasting gold price changes: Rolling and recursive neural network models. *Journal of Multinational financial management 18 (5)*(5), 477–487.

Paye, B. S. (2012). 'déjà vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics 106 (3)*(3), 527–546.

Shahzad, H., H. N. Duong, P. S. Kalev, and H. Singh (2014). Trading volume, realized volatility and jumps in the australian stock market. *Journal of International Financial Markets, Institutions and Money 31*, 414 – 430.

Thampanya, N., J. Wu, M. A. Nasir, and J. Liu (2020). Fundamental and behavioural determinants of stock return volatility in asean-5 countries. *Journal of International Financial Markets, Institutions and Money 65*, 101193.

Todorova, N. and M. Souček (2014). The impact of trading volume, number of trades and overnight returns on forecasting the daily realized range. *Economic modelling 36*, 332–340.

Vidal, A. and W. Kristjanpoller (2020). Gold volatility prediction using a cnn-lstm approach. *Expert Systems with Applications 157*, 113481.

Wang, X., C. Wu, and W. Xu (2015). Volatility forecasting: The role of lunch-break returns, overnight returns, trading volume and leverage effects. *International Journal of Forecasting 31 (3)*(3), 609–619.

Wang, Y.-H. and Y.-J. Hsiao (2010). The impact of non-trading periods on the measurement of volatility. *Review of Pacific Basin Financial Markets and Policies 13 (04)*(04), 607–620.

Wongbangpo, P. and S. C. Sharma (2002). Stock market and macroeconomic fundamental dynamic interactions: Asean-5 countries. *Journal of Asian Economics 13 (1)*(1), 27–51.

Xiong, R., E. P. Nichols, and Y. Shen (2015). Deep learning stock volatility with google domestic trends. *arXiv preprint arXiv:1512.04916*.

Xu, C. (2014). Expiration-day effects of stock and index futures and options in sweden: The return of the witches. *Journal of futures markets 34 (9)*(9), 868–882.

Xu, Y., D. Huang, F. Ma, and G. Qiao (2019). Liquidity and realized range-based volatility forecasting: Evidence from china. *Physica A: Statistical Mechanics and its Applications 525*, 1102–1113.

Zhu, X., H. Zhang, and M. Zhong (2017). Volatility forecasting using high frequency data: The role of after-hours information and leverage effects. *Resources Policy 54*, 58–70.