DOI: 10.20472/EFC.2023.019.021

AVRAHAM TURGEMAN
West University of Timisoara, Romania

CLAUDIU BOTOC
West University of Timisoara, Romania

MARILEN PIRTEA
West University of Timisoara, Romania

OCTAVIAN JUDE
West University of Timisoara, Romania

MODELLING INTRADAY REALIZED VOLATILITY: THE ROLE OF VIX, OIL AND GOLD

Abstract:

The main aim of the paper is to test an autoregressive implied volatility (IV) model that can significantly predict realized volatility (RV) of stock index. Subsequently, we want to test the predictive power of products that are external to the index of interest (S&P), by including certain commodities that are derived from VIX, i.e., crude oil and gold. The results do not reject the memory effect, given the predictive power of several lags for VIX over realized volatility. Furthermore, crude oil volatility is a significant predictor, alternatively in realized volatility and implied volatility. Finally, gold implied volatility (with higher lags) predicts stock returns volatility, which suggests a gap since traders tend to start gaining gold earlier to be on the safe side. Our findings have certain implications for trading and risk estimation.

Keywords:

Implied volatility, Realized volatility, AR model, Forecasting

JEL Classification: C22, G17

1. INTRODUCTION

The idea that volatility is a key concept in finance is supported by its extensive examination within empirical literature (by the date of writing this paper 6000+ papers in Web of Science include the topic “volatility forecasting.”). Volatility's role in financial markets is associated with core attributes like risk management, option pricing, and investing and therefore has implications for several stakeholders. The main characteristics of financial volatility consider persistence, clustering, and asymmetry. A conclusive description of estimating volatility can be found in D’Ecclesia and Clementi, (2021).

There are two generally accepted approaches to estimating and predicting volatility. One approach involves collecting information about future returns, from historical data, such as GARCH models or realized volatility. For example, GARCH models used to determine the volatility of returns deriving from financial assets have demonstrated a fine accuracy in estimating these volatilities. On the other hand, they have a reduced ability regarding estimation accuracy when these models are used to make forecasts — for example, the survey of Poon and Granger (2003). From a holistic point of view, we can say that GARCH models generate accurate in-sample estimates but less satisfactory out-of-sample forecasts. Another survey conclude that implied volatility provides more accurate forecasts than time-series models. According to their results, IV outperform RV in 76% of the studies considered, and GARCH in 94% of the studies considered (Poon and Granger, 2005)

Kambouroudis et al.(2016) carry out a comparative evaluation for the predictability of stock index volatility for several US & EU indices. In this respect, ten GARCH models and six ARMA models are considered for testing the hypothesis of volatility forecasts (GARCH, IV, and RV models). Overall, the results conclude that significant information for predicting future volatility is captured in all models, but with differences across markets. For US data the most significant model is the one that mixes information contained in ACGARCH forecasts with the information from options markets and information on RV from the ARFIMAX model. For EU data, the most significant model is the one that mixes asymmetric GARCH, IV, and RV through the ARMAX model.

Pilbeam et al. (2015) researched in their study whether different univariate GARCH models can forecast volatility better than implied volatility forecasts in the foreign exchange market. The data used in the study is segregated into two-time frames, first from 2002-2007, a period described as a low-volatility one, and from 2008 until 2012, a period featuring high volatility in the market. The authors conclude that none of the univariate GARCH models probed and researched has the forecasting ability of implied volatility, either on low or high-volatility periods. Implied volatility accommodates the data far more appropriately than the GARCH models. Another particular finding is that the GARCH models used perform particularly better in low-volatility periods than high ones.

The other approach involves the implied volatility of option pricing and highlights the possibility of using the most popular option pricing formula, i.e. the Black–Scholes (BS) equation, the volatility parameter being the only one unknown in this formula. Taking into account the linear relationship between volatility and option pricing, the BS formulas can be numerically inverted to derive an estimate for the volatility implied by the observed price option.

For example, Blair et al. (2010) make use of daily index returns and, sometimes, only intraday returns for the S&P100 index and the VIX and reveal that VIX provides forecasts with a higher level of accuracy when taking into consideration the increase in the forecast horizon than GARCH-type models. Investigating the low-frequency data after utilizing different ARCH models provides the foundation for the in-sample results, and uncovers no further evidence for incremental information in daily index returns other than that provided by the VIX index of implied volatilities. Out-of-sample forecasts indicate that VIX provides more accurate forecasts than both low-frequency and high-frequency index returns.

A consistent stream of literature that compares options-based forecasts with those from time series models can now be found, thanks to comprehensive research whose primary subject was the accuracy of volatility forecasting. It was linked to the launch of the VIX by the Chicago Board Exchange (CBOE).

Forecasting future volatility of different asset returns is of significant consideration for most market participations, for investment adjustments regarding derivative pricing and risk management. Wang et al (2016) analyze in their study the intraday VIX of the CBOE, searching for the best timing for gathering relevant information for predicting realized volatility. Their findings show that the period before the market closes, at noon US time. The high forecasting output levels found around noon suggest that there is less complex and intricate trading motivation in that time frame, thus the VIX comprises greater informative market data of future volatility. For a similar methodology, Kambouroudis et al. (2021) find that IV is a robust predictor for volatility forecasting. Furthermore, the HAR-IV model is superior to HAR without IV, with certain particularities for different markets, i.e. leverage effect, overnight returns, and volatility of realized volatility respectively.

Kourtis et al. (2016) examine the predictive output and economic significance of implied, realized, and GARCH volatility models within an international portfolio framework. They employ several models and thus can compare them with the support of different statistical tests and loss functions. The main findings suggest that the Heterogeneous Autoregressive (HAR) model of Corsi (2009) provides the most accurate results for deriving 1-day ahead forecasts, i.e. incorporating volatility risk premium is more appropriate for predicting.

At the same time, standard theory suggests that realized volatility is a powerful tool for the measurement, modeling, and forecasting of high-frequency data (Andersen et al. 2003; Andersen et al. 2004). Frijns and Margaritis (2008) assess to what extent intraday data can explain and predict end-of-the-day volatility. They find that the explanatory power of first-hour volatility for daily volatility is as high as 68%, whereas the average volatility generated during this first hour is <30%.

Driesprong et al. (2008) evidence that for the in-sample approach variations in oil prices predict stock returns whereas Chen et al. (2010) evidence, both in-sample, and out-of-sample, only unidirectional relationship between currency and global commodity prices. Wang et al. (2018) extend both previous empirical works, intending to predict stock volatility with oil volatility. On the one hand, they find evidence for their hypothesis for both in-sample and out-of-sample approaches. On the other hand, they find that crude oil volatility exhibits predictive power over stock return volatility.

Dai et al. (2020) use the AR benchmark model from Wang et al. (2018) and further extend it by including implied volatility as an additional predictor for stock realized volatility. The in-sample results suggest Granger causality from implied volatility to stock volatility. Further, the out-of-sample results suggest that for stock markets implied volatility improves the predictability of volatility.

Given these relative findings, we aim to participate in the ongoing debate of predictability for clearer findings. In this respect, we aim to test an autoregressive implied volatility model that can significantly predict realized volatility (RV) of stock index. Subsequently, we want to test the predictive power of products that are external to the index of interest (S&P), by including certain commodities that are derived from VIX, i.e., crude oil and gold. Several contributions can be made. First, we want to test if other commodities apart from crude oil (i.e., gold) have predictive power over RV. Second, unlike Dai et al. (2020), we test a multivariate AR model, by including significant lags for independent variables. Third, we examine a recent intraday sample (15th January – 15th April 2022), with 5 minutes timeframe, which allows controlling for current macroeconomic and geopolitics conditions. The results do not reject the memory effect, given the predictive power of several lags for VIX over realized volatility. Furthermore, crude oil volatility is a significant predictor, alternatively in realized volatility and implied volatility. Finally, gold implied volatility (with higher lags) predicts stock returns volatility, which suggests a gap since traders tend to start gaining gold earlier to be on the safe side.

The paper is structured as follows. After the introduction in section 1, in section 2 we define the data and methodology of the research. In section 3 we perform an analysis of the data, using a few common time series tests, based on statistical methods and finally, we fit the model and test its performance. Section 4 checks the sensitivity of our analysis whilst in section 5 we conclude the findings of the research and its implications.

2. DATA AND METHODOLOGY

We use the S&P realized volatility as the variable of interest and we use implied volatility (represented by the VIX index) as the main independent variable. Further, we include Crude Oil and Gold indices as the controls. The data are collected from the Refinitiv database, for the period between 15th January – 15th April 2022 (maximum length available of 5000 observations for each variable), 5 minutes time frame. We define the Realized volatility as the standard deviation of the samples in a given time frame which is calculated as follows:

$$R_i \triangleq \log P_i - \log P_{i-1} \qquad (1)$$

$$\overline{R_t} \triangleq \frac{1}{\Delta t} \cdot \sum_{i=t-\Delta t}^{t} R_i \qquad (2)$$

$$S\&P\ RV \triangleq RV_t = \sqrt{\frac{\sum_{i=t-\Delta t}^{t} (R_i - \overline{R_t})^2}{\Delta t - 1}} \qquad (3)$$

Where:

P_i - the assets price at time i

R_i - the return at time i

$\overline{R_t}$ - the average return over time frame t-Δt

RV_t - realized volatility at time frame t (the standard deviation of the returns)

Δt = the number of samples to accumulate, equal to 12 for our case, which corresponds to 60 minutes of trading.

Furthermore, we also define the following independent variables, that are derived from the VIX, Crude Oil, OVZ (Crude Oil Implied Volatility), Gold, and GVZ (Gold Implied Volatility):

$Crude\ RV_t, OVZ\ RV_t, GVZ\ RV_t$ – the realized volatility of the assets, calculated the same way as the S&P realized volatility, which means that it is the standard deviation of the asset's returns.

We also define the following variables that are derived from the base variable:

$$OVZ\ LG_t = \log P_t - \log P_{t-1} \qquad (4)$$

It is worth to be mentioned that $OVZ\ LG_t$ is a non-cumulative derivative of the $OVZ\ RV_t$. Similar formulas are used for $VIX\ LG_t$ and $GVZ\ LG_t$.

The summary statistics of all variables considered are reported in Table 1, whereas can notice that the mean range is between -7.4E-05 to 0.000251 and is highly persistent, in line with previous findings (Christiansen et al., 2012; Nonejad, 2017). Minimum and maximum values could suggest higher persistence during the opening and closing times of different international financial markets (Serknas, 2013).

Table 1: Summary statistics

STATS	SP	VIX	CRUDE	GOLD	OVZ	GVZ
MEAN	-7.4E-05	2.77E-06	-8.96E-06	1.33E-05	0.000251	0.000165
MEDIAN	0.0000	-0.0008	0.0002	0.0000	-0.0008	0.0000
ST. DEV.	0.0023	0.0156	0.0039	0.0031	0.0122	0.0098
MIN	-0.0124	-0.0562	-0.0308	-0.1014	-0.0805	-0.1083
MAX	0.0098	0.1130	0.0198	0.0591	0.0866	0.0852
SKEWNESS	-0.2086	1.0968	-0.9965	-5.7925	1.4132	0.0871
KURTOSIS	6.2540	7.9146	8.8445	346.8077	12.6929	18.3671

In terms of methodology, Dai et al. (2020) demonstrate the autoregressive nature of the intraday realized volatility. In this paper we test the auto-regression nature of the intraday volatilities, using more indicators. After showing a high probability for auto-regression possibility, we use the Vector Auto Regression (VAR) model which allows the use of multivariate AR. The mathematic representation of the model is:

$$Y_t = \mu_0 + \sum_{j=1}^{p} \alpha_j \cdot Y_{t-j} + \sum_{i=1}^{n} \sum_{j=1}^{p} \beta_{ij} \cdot X_{i_{t-j}} + \varepsilon_t \qquad (5)$$

Where:

Y_t - the dependent variable that is being predicted

p - the lags in time that correlate to Y_t

X_i - the independent variables

α, β - the coefficients of the variables

ε_t – sample error

3. RESULTS

3.1. Stationarity Test

A basic assumption in AR models is that the signals are stationary. It means that their average and variance are constant in time. To test stationarity, we use the Augmented Dickey-Fuller test. We use a null hypothesis that the signal is not stationary and we try to reject it with sufficient significance (usually 5%).

The following table summarizes the ADF test results for all variables considered:

Table 2: ADF results

	SP RV	**Crude RV**	**VIX**	**OVZ**	**GVZ**	**VIX LG**	**OVZ LG**	**GVZ LG**	**VIX RV**	**OVZ RV**	**GVZ RV**
ADF	-8.0	-11	-2.2	-1.7	-1.7	-13	-27	-53	-9.6	-8.9	-11
Critical ADF (5%)	-2.9	-2.9	-2.9	-2.9	-2.9	-2.9	-2.9	-2.9	-2.9	-2.9	-2.9
p-val	0.00	0.00	0.2	0.43	0.45	0.00	0.00	0.00	0.00	0.00	0.00
Stationary	**Yes**	**Yes**	No	No	No	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

According to the ADF test of stationarity, we use in the model only the signals that are significantly stationary at a 5% level (p-val<0.05). In the first phase, we continue with only stationary ones, although non-stationary ones can also be proven useful, especially in short-time prediction.

3.2. Granger Causality test

Granger Causality is a statistical test that indicates whether there is a correlation between one signal and another lagged signal. If a signal Granger causes another, it means that there is a good probability that. can help predict it. The following table summarizes the results of the Granger Causality test of the different variables with respect to the predicted variable S&P RV. For the lags of each variable, we calculate the p-value, and we expect to find significant results at 10%, which means the p-value<0.1. We first try to find Granger Causality in the first 3 lags, and for variables that do not have significant GC in the first 3 lags, we test again with a maximum of 15 lags and try to find a lag with significant GC.

Table 3: Granger test results

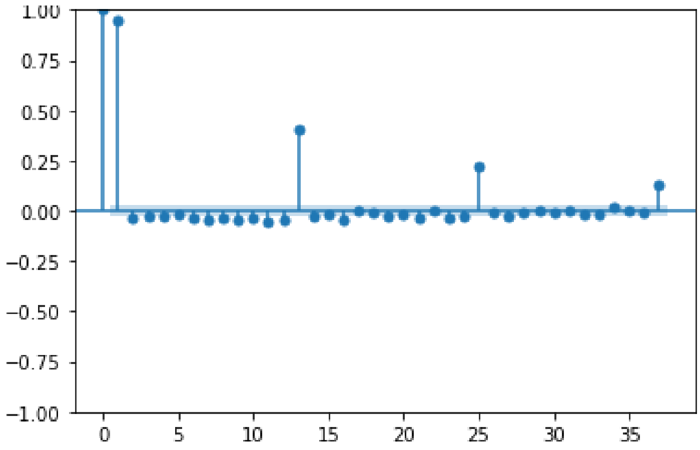
Specification	**Crude RV**	**VIX LG**	**OVZ LG**	**GVZ LG**	**VIX RV**	**OVZ RV**	**GVZ RV**
Lag 1	0.6	0.01	0.9	0.5	0.3	0.03	0.2
Lag 2	0.6	0.07	0.04	0.5	0.2	0.1	0.3
Lag 3	0.8	0.15	0.2	0.7	0.08	0.18	0.2
First significant lag	--			12			7

The conclusion is that Crude RV does not Granger Cause S&P RV up to 15 lags, while the rest of the variables do. Gold signals seem to lag pretty far behind the S&P RV, with GVZ RV having significant GC in lags 7-15.

3.3. PACF Test of S&P RV

The PACF (Partial Auto Correlation Function) test shows the autoregressive nature of a signal, by testing the correlation of a signal with its lagged self. If a signal is correlated with its lagged self, It means that there are significant α_j coefficients in the model. The following is a plot of the S&P RV PACF chart:

Figure 1: Partial Auto Correlation Function - SPRV



The PACF chart indicates significant coefficients in lags=1,13,25,37. To keep the model simple and significant, we use only lags 1 and 13.

3.4. Fitting The Model

We split the data into training data and testing data. The training data is used to train the model and get the coefficients while testing data is used to test the model out of samples. The training data contains 80% of the samples and the testing data contains the rest. To find the optimal lags for the VAR model, we use the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Akaike's Final Prediction Error Criterion (FPE), and the Hannan-Quinn information criterion (HQIC). According to the preemptive tests, we expect the model to be most efficient at lags=13 and have few significant coefficients.

The following list is the calculation of the criteria of the model in different lags. The model is expected to be optimal when the different criteria are minimal:

Table 4: VAR order selection

#Lags	AIC	BIC	FPE	HQIC
0	-59.62	-59.61	1.28E-26	-59.62
1	-63.83	**-63.78***	1.91E-28	-63.81
2	-63.84	-63.75	1.88E-28	-63.81
3	-63.84	-63.72	1.88E-28	-63.8
4	-63.85	-63.68	1.87E-28	-63.79

5	-63.84	-63.64	1.88E-28	-63.77
6	-63.84	-63.6	1.88E-28	-63.75
7	-63.84	-63.56	1.88E-28	-63.74
8	-63.84	-63.52	1.88E-28	-63.72
9	-63.84	-63.48	1.88E-28	-63.71
10	-63.84	-63.44	1.88E-28	-63.7
11	-63.85	-63.41	1.86E-28	-63.69
12	-63.89	-63.41	1.80E-28	-63.72
13	**-64.16***	-63.64	**1.365e-28***	**-63.98***
14	-64.15	-63.59	1.37E-28	-63.96
15	-64.15	-63.55	1.38E-28	-63.94

We notice that AIC, FPE, and HQIC point to lags=13, while BIC points to lag=1. The difference is mainly because BIC penalizes higher orders more than the other criteria. Since the results of the previous tests of Granger Causality showed significant values in higher order than lag=1, and since the PACF test of the S&P RV showed significant autoregressive coefficients at lag=13, we prefer to follow the AIC, FPE, and HQIC criteria and pick lags=13. Thus, we fit the VAR model with lags = 13 using the training data, and we get significant coefficients.

3.5. In Sample Evaluation

In this step, we test the model results by predicting the volatility of the in-sample data, meaning the training data, given that in-sample predictability is a pre-condition for out-of-sample predictability (Inoue and Kilian, 2005). We evaluate the model according to two main factors – the Mean Squared Error (MSE) and the coefficient of determination (R-Squared or R^2). A good model should have a low MSE and high R^2 factor (Brooks, 2019). MSE and R^2 definitions have the following formulas:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \qquad (6)$$

whereas R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (7)$$

where:

y_i – the actual value of sample i

$\tilde{y}_i$ – the predicted value of sample i

$\bar{y}$ – mean value of all the samples

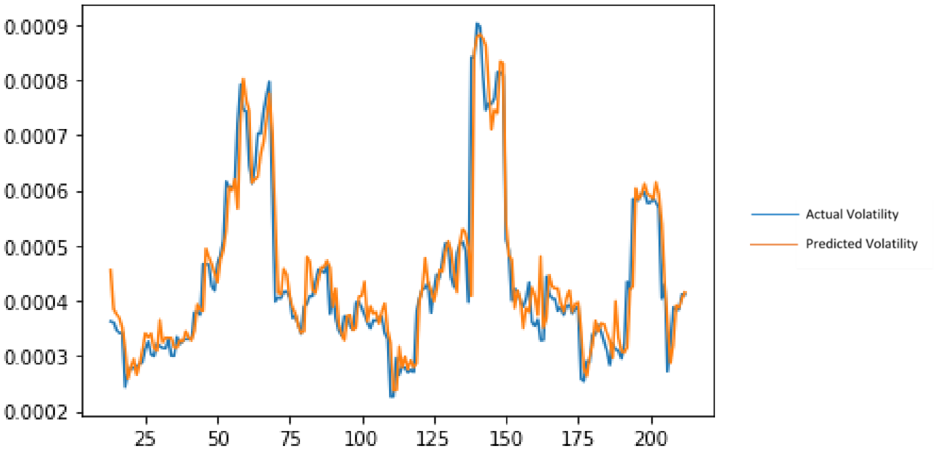
We expect to observe that the MSE decreases and R^2 increases as we add more lags to the model until we reach lags = 13, which is the optimum according to the model's criteria. The results show the expected behavior, and the model evaluation with lags = 13 (MSE = 1.2e^{-8}; R^{2} = 0.92) yields the following results:

Table 5: In-sample regression results for Eq. (6) with 13 lags – significant coefficients

	Coefficient	Std. Error
CONST	0.000026***	0.000004
L1.SP RV	0.985135***	1.54E-02
L1.VIX LG	0.000641*	3.90E-04
L2.Crude IV LG	7.39E-04**	0.000356
L6.Crude IV LG	0.000621*	0.000363
L7.Crude IV LG	0.001081***	0.000363
L9.VIX LG	0.000647*	0.000393
L10.Crude RV	0.003005**	0.001489
L11.Crude RV	-0.00383***	0.001491
L11.VIX LG	0.000732*	0.000393
L12.SP RV	-0.39193***	0.020969
L12.VIX LG	-0.0012***	0.000394
L12.Crude IV LG	-0.00182***	0.000356
L12.Gold IV LG	0.003279***	0.000586
L13.SP RV	0.369213***	0.015006
L13.Gold IV LG	0.00132**	0.000581

The coefficient estimates of α and βs are significant at a 1% level for the SP, VIX, Crude, and GOLD which indicates the strong in-sample predictability from stock and commodity market implied volatility to stock volatility for the lag order p = 13. We find that SP has a statistically negative response to most signals starting with the eleventh lag, including its own lagged signal at lag=12. For Gold one can report significant predictors for p=12 and p=13, which suggests a delay of 60 minutes, consistent with its property as a safe haven (Jubinski and Lipton, 2013). The following chart shows the results for the volatility forecasting performance of a randomly selected 200 samples, whereas mainly actual volatility overlaps with predicted volatility.

Figure 2: SP Realized Volatility: Predicted vs. Actual



3.6. Out-of-Sample Evaluation

We test two methods for out-of-sample prediction. In the first method (Method A) we train the model and use constant coefficients. Then we use a rolling window over the testing data for prediction. In the second method (Method B), we use a rolling window for both fitting and predicting. Prediction is always made 1 step forward into the future.

Method A is faster and more efficient because the fitting is done only once and then the coefficients are constant. In Method B, however, fitting is done on each step which makes it less efficient but might be more accurate. Method B might not be possible to implement in real-time when the sample rate is in the milliseconds or microseconds due to relatively longer processing time, so it's important to test it and compare it to method A and see if there's any accuracy benefit.

Both methods are sensitive to the size of the rolling window which affects the accuracy of the model so we compare the two methods on different sizes of rolling windows.

The way to predict the new data in time series is by predicting only 1 sample forward in time and then updating the input data to the model with the new known sample. Meaning, we use the last available "known" samples to predict the next sample.

Going over the test data, we get the following results. In method B "size" means the size of the rolling window frame in samples, in method A "size" means the size (in samples) of the training data that was used to fit the model once:

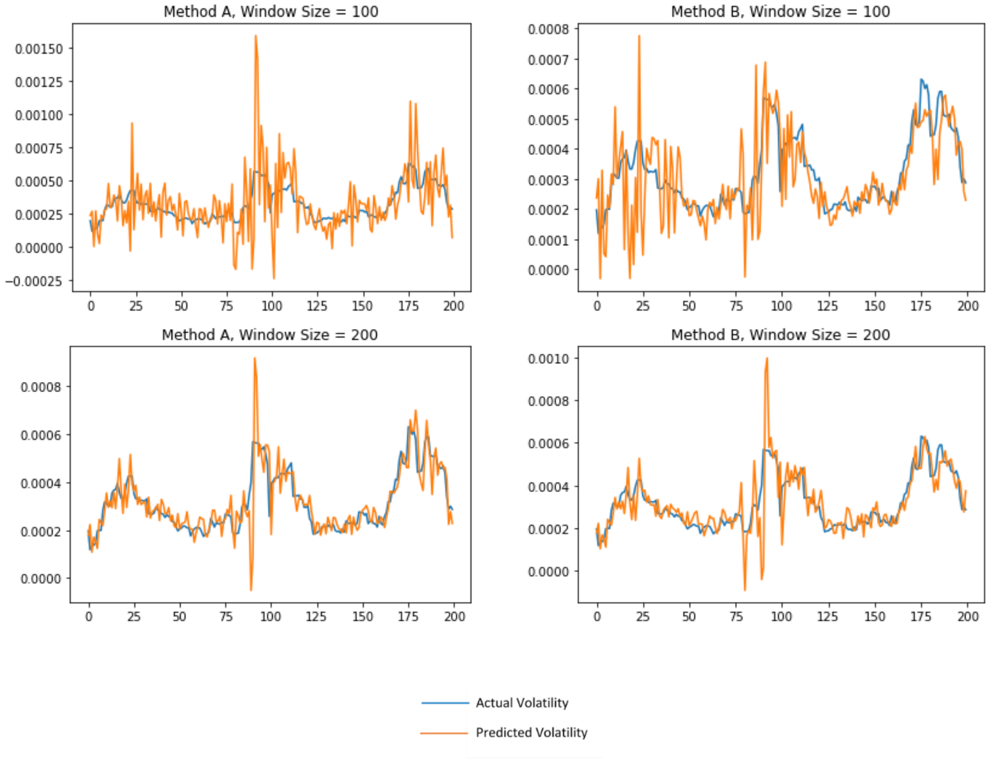
Table 6: Out-of-sample results

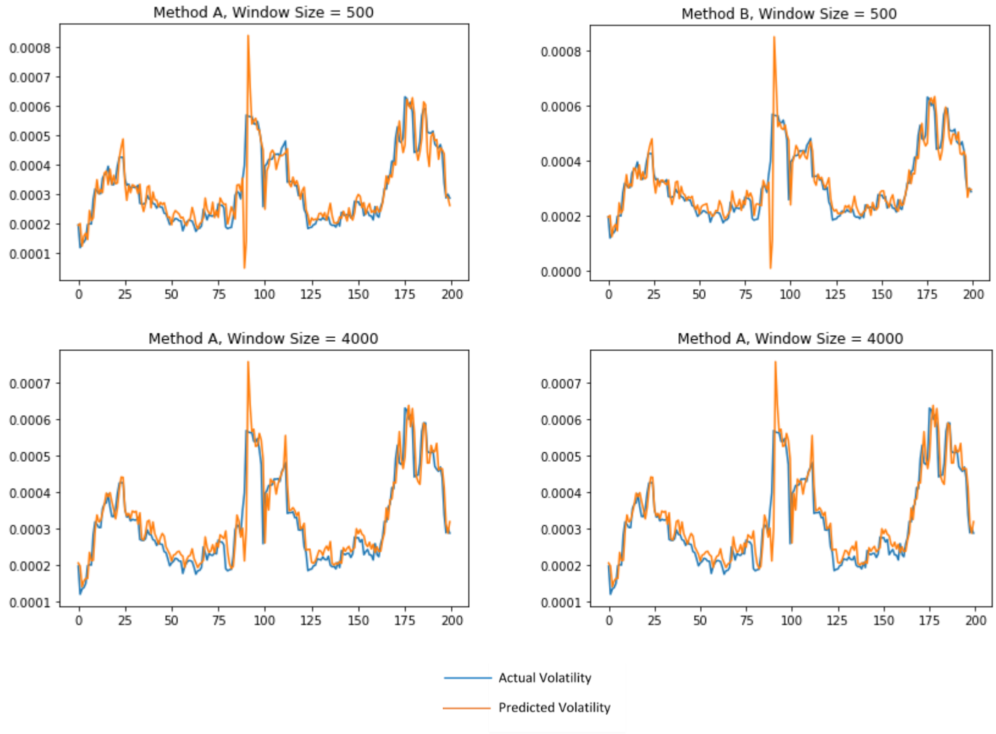
Size	**100**	**200**	**500**	**4000**
Method A				
MSE	5.6e^{-8}	9.0e^{-9}	5.5e^{-9}	4.1e^{-9}
R^2	-0.42	0.77	0.86	0.90
Method B				
MSE	2.2e^{-8}	7.7e^{-9}	5.1e^{-9}	4.1e^{-9}
R^2	0.47	0.81	0.88	0.90

Please note that negative R^2 means that the model accuracy is worse than a straight line of a constant value.

The following chart shows the results of randomly selected samples for each of the methods Overall there are 5000 samples. We use up to 4000 samples for training the model (that's the different window size), and then we test it on the next 1000 samples:

Figure 3: SP Rolling Realized Volatility: Predicted vs. Actual





We notice that when the rolling window is too small (less than 200 samples), both methods fail to predict the volatility. However, when starting to increase the number of training samples to above 200 samples, there is a clear advantage to predictions using method B, until the rolling window reaches 4000 samples and then both methods show equal prediction accuracy.

The conclusion is that a model's accuracy decreases with time and then there is a need to fit the model again with fresh data. If the model is fitted more frequently, then fewer samples can be used. However, there is a limit to this conclusion, because if the training data window size is too small, the model fails to predict even one step forward.

This result is important because it means that for intraday real-time trading, it is possible to fit a VAR model dynamically which makes it unnecessary to train the model in advance since it can be trained in real time over a relatively smaller-sized window frame.

3.7. Analyzing the Residuals

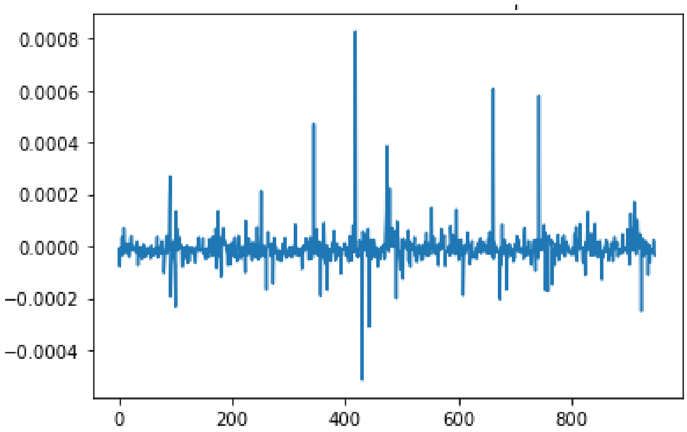
An important part of evaluating the model is evaluating its residual errors. In a good model, we expect the residual error to be as close as possible to white noise, which means that there is no more modeling left to do since the error is purely random. We test the residuals for the case of method B with 4000 samples rolling window size.

We define the residual error as:

$$\varepsilon_t = y_t - \tilde{y}_t \qquad (8)$$

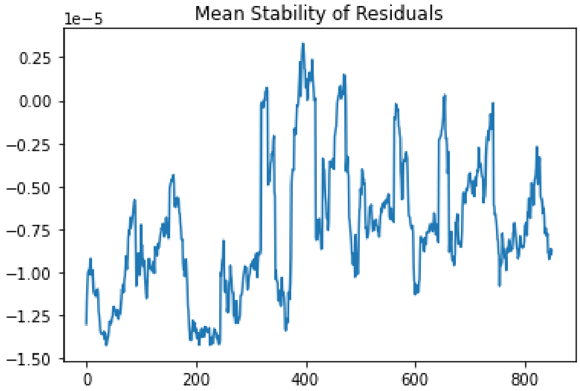
Mean value: The mean value of the white noise should be 0. In our case, we get $\bar{\varepsilon} = 7.7 \cdot 10^{-6}$. The following chart shows the residual error over time and we test it according to the following criteria.

Figure 4: Residual error plot



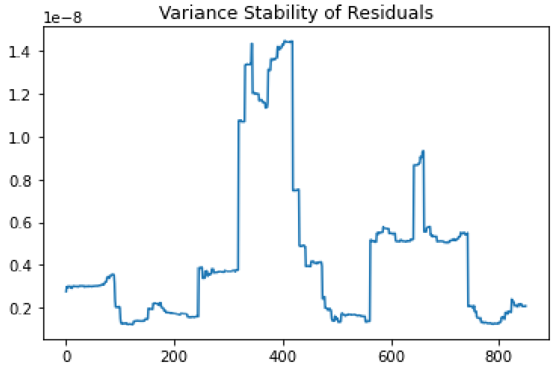
Mean Stability over time: The mean value of white noise is constant over time. The following chart shows the mean of the residual error for every 100 samples. The mean is very low, but it still has some seasonality in it which means that the residual error is not pure white noise.

Figure 5: Mean stability of residuals



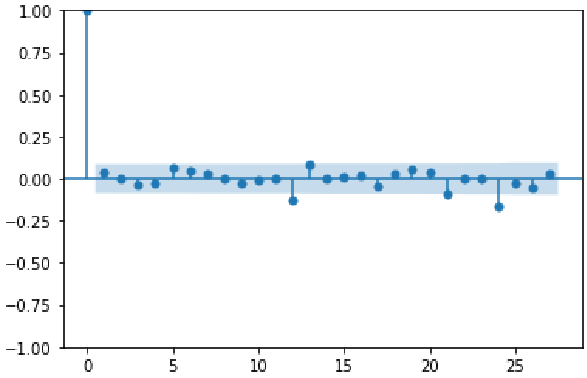
Variance stability over time: The variance of white noise is constant over time. The following chart shows the variance of the residual error every 100 samples. Although the variance is very low, we get some changes in the variance over time.

Figure 6: Variance stability of residuals



Auto Correlation The autocorrelation of white noise is zero, and we test the residual error accordingly:

Figure 7: Auto Correlation of residuals



We notice that few lags are on the edge of significance, which means that the signal's autocorrelation is not pure zero.

Looking at the few white noise criteria we conclude that the residual error is not pure white noise which means that the model can still be improved, although the prediction accuracy is pretty good. We suspect that since we model intraday data, then the gap between the days might be the cause for the seasonality.

4. ROBUSTNESS TESTS

As noted in section 2 – Data and Methodology, in this paper we choose Δt=12 for calculating the realized volatility. Since the time frame is for 5 minutes, by choosing Δt =12 we calculate the realized volatility in a time frame of 1 hour.

However, to test the robustness of the findings, as Rossi and Inoue (2012) suggest the issue of the estimation window, we run the same tests with Δt =9 which is equal to a time frame of 45 minutes, and Δt = 6 which equals to a time frame of 30 minutes.

When testing for shorter time frames, we expect the models to be less accurate since the data will have a worse signal-to-noise ratio. Averaging a larger number of samples is expected to result in smoother data and therefore more significant results. Thus, it is interesting to test whether a shorter time frame can still yield a significant model because the actual implication for such results is faster response time to the market's movements.

For the robustness tests, we show the out-of-sample results of only one of the methods that were proven successful for Δt =12, assuming that an unsuccessful method for Δt =12 will only get worse for Δt <12.

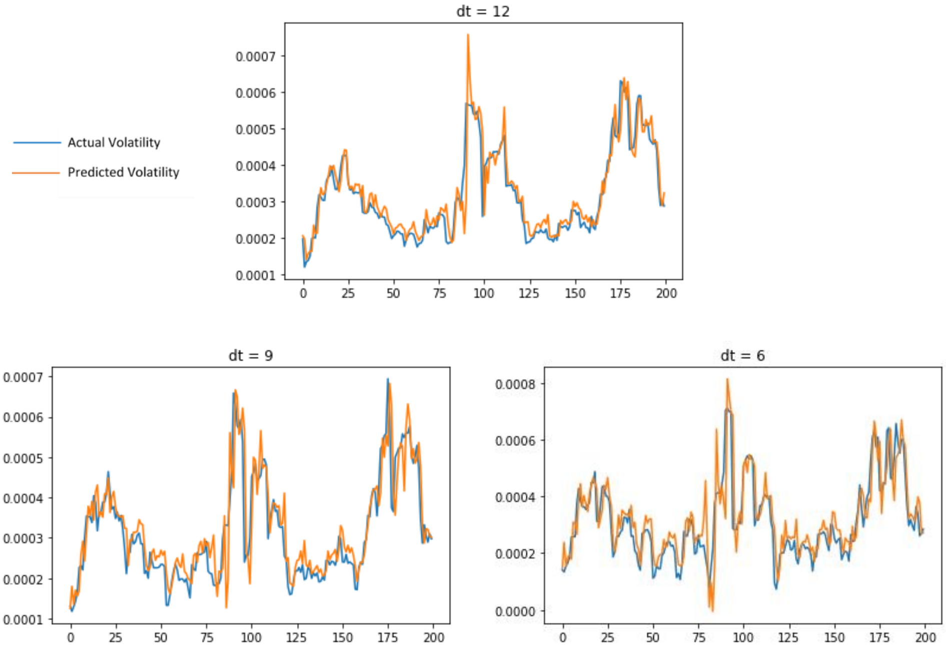
First, we analyzed the MSE and R^2 parameters of the model, which is displayed in the following table:

Table 7: Out-of-sample results Robustness results for size 4000 and Δt = 6, 9, 12

MSE	4.1e^{-9}	6.2e^{-9}	1.1e^{-8}
R^2	0.90	0.86	0.78

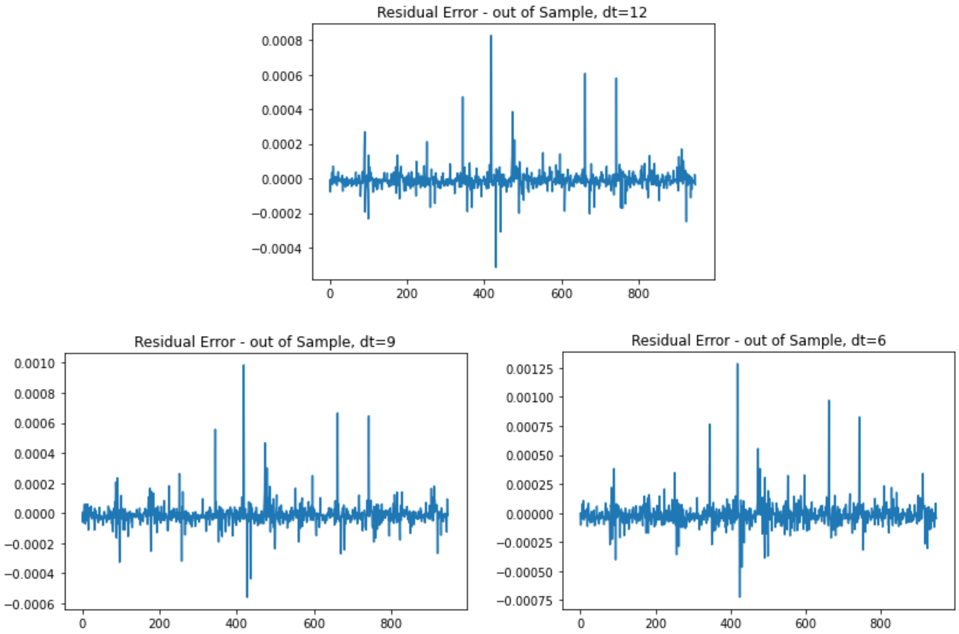
We see that as expected, the R^2 is lower when the t is lower. However, the MSE factor is about the same. The following charts show the actual volatility versus the predicted volatility:

Figure 8: SP RV. The blue line is the actual volatility. Orange is the predicted volatility



The increase in noise and thus lower signal-to-noise ratio when the t is smaller can be noticed easily in the charts. To complete the robustness test, we also compared the residual error of the model with the different t factor:

Figure 9: Residual error plot - robustness



We can see that the residual error for out-of-sample data has some higher peaks when t is lower, but even the t=6 model does not seem to diverge.

When analyzing the robustness test results, we see that the results are as expected and fit the theory pretty well. When using a lower t, which means that we averaging a smaller number of samples, we can see that although there is an expected decrease in the accuracy of the model, the model remains significant and therefore useable. The benefit of a lower t is a better response to market movements and therefore, using a lower t is expected to have its advantages when practicing actual trading.

5. CONCLUSIONS

This paper investigates the connection between the volatility of several financial and commodities products and tries to find a model that can significantly predict the intraday realized volatility of the S&P 500 stock index.

We find significant predictability of the S&P 500 among VIX, Crude, and Gold. Interestingly, for Crude, we found the predictability in both the realized and implied volatility while in others we find it only in the implied volatility. Using a few common tests, this paper shows that the autoregressive nature of the realized volatility which was demonstrated for daily returns also valid for intraday returns

Those findings have certain implications for trading and risk estimation. First, the autoregressive nature of the realized volatility is persistent during intraday trading which allows efficient earlier risk-related alerts. In other words, such models can produce significant profits since statistical accuracy can coincide with trading profitability. Second, there is a significant predictive power of the S&P 500 in other products which are external to the S&P 500 index, such as Oil and Gold.

REFERENCES

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

Andersen, T. G., Bollerslev, T., & Meddahi, N. (2004). Analytical evaluation of volatility forecasts. *International Economic Review*, *45*(4), 1079-1110.

Blair, B. J., Poon, S. H., & Taylor, S. J. (2010). Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. In *Handbook of quantitative finance and risk management* (pp. 1333-1344). Springer, Boston, MA.

Brooks, C. (2019). *Introductory Econometrics for Finance 4E*. Cambridge University Press.

Chen, Y. C., Rogoff, K. S., & Rossi, B. (2010). Can exchange rates forecast commodity prices? *The Quarterly Journal of Economics*, *125*(3), 1145-1194.

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

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

Dai, Z., Zhou, H., Wen, F., & He, S. (2020). Efficient predictability of stock return volatility: The role of stock market implied volatility. *The North American Journal of Economics and Finance*, *52*, 101174.

Driesprong, G., Jacobsen, B., & Maat, B. (2008). Striking oil: another puzzle?. *Journal of financial economics*, *89*(2), 307-327.

D'Ecclesia, R. L., & Clementi, D. (2021). Volatility in the stock market: ANN versus parametric models. *Annals of Operations Research*, *299*(1), 1101-1127.

Frijns, B., & Margaritis, D. (2008). Forecasting daily volatility with intraday data. *The European Journal of Finance*, *14*(6), 523-540.

Kambouroudis, D. S., McMillan, D. G., & Tsakou, K. (2016). Forecasting stock return volatility: A comparison of GARCH, implied volatility, and realized volatility models. *Journal of Futures Markets*, *36*(12), 1127-1163.

Kambouroudis, D. S., McMillan, D. G., & Tsakou, K. (2021). Forecasting realized volatility: The role of implied volatility, leverage effect, overnight returns, and volatility of realized volatility. *Journal of Futures Markets*, *41*(10), 1618-1639.

Kourtis, A., Markellos, R. N., & Symeonidis, L. (2016). An international comparison of implied, realized, and GARCH volatility forecasts. *Journal of Futures Markets*, *36*(12), 1164-1193.

Inoue, A., & Kilian, L. (2005). In-sample or out-of-sample tests of predictability: Which one should we use?. *Econometric Reviews*, *23*(4), 371-402.

Jubinski, D., & Lipton, A. F. (2013). VIX, gold, silver, and oil: how do commodities react to financial market volatility?. *Journal of Accounting and Finance*, *13*(1), 70-88.

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.

Pilbeam, K., & Langeland, K. N. (2015). Forecasting exchange rate volatility: GARCH models versus implied volatility forecasts. *International Economics and Economic Policy*, *12*(1), 127-142.

Poon, S. H., & Granger, C. W. (2003). Forecasting volatility in financial markets: A review. *Journal of economic literature*, *41*(2), 478-539.

Poon, S. H., & Granger, C. (2005). Practical issues in forecasting volatility. *Financial analysts journal*, *61*(1), 45-56.

Rossi, B., & Inoue, A. (2012). Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economic Statistics*, *30*(3), 432-453.

Serknas, D. (2013). Time Series Model for Forecasting Intraday Volatilities.

Wang, Y., Ma, F., Wei, Y., & Wu, C. (2016). Forecasting realized volatility in a changing world: A dynamic model averaging approach. *Journal of Banking & Finance*, *64*, 136-149.

Wang, Y., Wei, Y., Wu, C., & Yin, L. (2018). Oil and the short-term predictability of stock return volatility. *Journal of Empirical Finance*, *47*, 90-104.