Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/45/20230305

High-Frequency Data Analysis for Volatility Prediction of the CSI 300 ETF: An Empirical Research Using HAR Family Models

Xiao Lin[1,a,*]

[1]*School of Finance, Zhejiang University of Finance & Economics, Hangzhou, 310018, China*
a. 210504011011@zufe.edu.cn
**corresponding author*

Abstract: In this paper, four HAR family models are used to study them and five-minute high-frequency trading data of the CSI 300 ETF is the target. Descriptive statistics is firstly conducted, and it is found that all variables have obvious autocorrelation and long-term memory which are suitable for time series analysis. Then, in-sample and out-of-sample analyses are implemented to test the predictive effect of each component on the volatility of the CSI 300 ETF and to compare the predictive ability of each model. The empirical results of the in-sample data show that daily and weekly realized volatility, continuous volatility, daily and monthly jump volatility, daily and monthly realized positive semi-variance, weekly and monthly realized negative semi-variance and positive and negative signed jump variation have stronger predictive effects on volatility of the CSI 300 ETF, while weekly realized volatility, jump volatility and realized positive semi-variance, and daily realized negative semi-variance have a weaker predictive effect on the CSI 300 ETF volatility. The results of the MCS test for out-of-sample prediction show that the HAR-RV-CJ model and HAR-RV-RSV model have significantly better out-of-sample predictive ability on the CSI 300 ETF volatility than the other two HAR family models, with HAR-RV-RSV model exhibiting the highest predictive accuracy in most cases. The above results indicate that the jump factor and the asymmetric factor of positive and negative returns can effectively improve the forecasting accuracy of HAR family models, so these two factors cannot be ignored in the construction of future HAR family models.

Keywords: HAR-RV model, rolling window prediction, MCS test

1. Introduction

The study of volatility in financial markets is the basis for scholars to study practical financial issues such as asset allocation, option pricing, and risk management. Accurate measurement and forecasting of the volatility of the financial markets are necessary for a quantitative understanding of these practical challenges. As a result, scholars have been interested in that of the financial markets.

On April 8, 2005, Shanghai and Shenzhen Stock Exchanges released the CSI 300 Stock Index. The constituent stocks of the index are selected based on two fundamental criteria: size and liquidity, covering nearly 60% of the market's capitalization. So far, the movements of the CSI 300 Index are still used by institutions as a benchmark for evaluating investment returns, representing the basic

DOI: 10.54254/2754-1169/45/20230305

situation of mainstream investment in the market and reflecting the basic operation of the Chinese financial market. Since the Index cannot be traded, the CSI 300 ETF is chosen as the research object in this paper. The CSI 300 ETF is an open-ended index fund that is traded and redeemed in the secondary market based on the CSI 300 index, which can reflect the fluctuations of the Chinese financial market better than others because it can be traded directly.

In the past two decades, with the reduction of data storage costs and the continuous improvement of financial databases, it has become increasingly easy to obtain intraday high-frequency data in financial markets, which makes financial high-frequency data an increasingly important tool for studying financial market volatility. In 1998, Realized volatility, which Andersen and Bollerslev first proposed based on high-frequency trading data, is a novel volatility measurement [1]. Andersen et al. studied the U.S. foreign exchange market and showed that realized volatility had significant long-term memory and realized volatility estimators did not depend on any model, having no estimated error, and can be used as an unbiased and effective estimator of real volatility [2]. Realized volatility has been used to study market characteristics such as autocorrelation, aggregation, asymmetry, and long-term memory. Then, given the heterogeneous market hypothesis, Corsi proposed a heterogeneous autoregressive realized volatility model, which divides the volatility of financial markets into three distinct volatility components [3]. According to empirical findings, the HAR-RV model has significantly higher predictive power than the GARCH model for the future volatility of financial markets. Although this model is not similar to traditional autoregressive models such as ARCH, it can be estimated by ordinary least squares, which is its unique advantage, and it can also use the economic implications of images to portray thick-tailed characteristics and long-term memory.

Many academics built new volatility forecasting models based on the benchmark model after developing the fundamental HAR-RV model in an effort to increase the predicted accuracy of the enhanced HAR model for financial market volatility. The most notable among them is Andersen, who built the HAR-RV-CJ model based on HAR-RV, which significantly improved the model's prediction. Andersen' finding was based on the quadratic variance theory proposed by Barndorff-Nielsen and Shephard, which decomposed the realized volatility into continuous path variance and discrete jump variance [4, 5]. Barndorff-Nielsen split realized volatility into positive and negative semi-variances [6]. Patton and Sheppard subtracted the latter one from the former one to obtain the signed jump variance and Lyocsa and Stasek employed a full-subset regression approach in conjunction with the HAR model and discovered that the HAR-CSQR model outperformed the benchmark HAR model for four significant market indices [7, 8]. Hong and Hwang considered a multivariate HAR model with index weighted and applied it to U.S. stock prices and found that the error of that model was smaller than the traditional model [9]. Based on the reasearches of the previous study, this paper suggests using the four HAR families' most traditional models—HAR-RV, HAR-RV-CJ, HAR-RV-RSV, and HAR-RV-SJPN—to see if they can predict financial market volatility more accurately.

2. Methodology

2.1. Data Selection

The data source is the Wind database's 5-minute high-frequency data of the CSI 300 ETF, with a sample period from January 1, 2013, to December 31, 2022, with a total of 2430 trading days, omitting holidays and missing data. China's Shanghai (Shenzhen) Exchange normal business hours are 9:30 - 11:30 and 13:00 - 15:00 for a total of 4 hours (240 minutes) of trading time, divided by 5 minutes as the unit of trading time, thus there will be 48 valid data per day, data capture including trading hours, closing prices, volume, turnover, a total of 116,640 groups of high-frequency trading data. The data processing and the subsequent regression empirical evidence are used in R language.

DOI: 10.54254/2754-1169/45/20230305

2.2. Variable Construction

2.2.1. Realized Volatility

Realized volatility, defined as the sum of the squares of the observable high-frequency returns over the course of a trading day, was introduced by Andersen and Bollerslev [1]. It is calculated as follows: assume a trading day t and split the day's trading into N segments, P_{ti} stands for the ith closing price during the trading day of t., $i = 1,2,\cdots,N$. Let $r_{t,i}$ be the logarithmic return of the ith segment in trading day t, $r_{t,i} = 100 * (lnP_{ti} - lnP_{t,i-1})$ The trading day's daily realized volatility RV_t^d can be represented as follows:

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

The computation of weekly and monthly realized volatility by Corsi was then added to that of daily realized volatility [3].

$$RV_t^w = (RV_t^d + RV_{t-1}^d + \cdots + RV_{t-4}^d)/5 \tag{2}$$

$$RV_t^m = (RV_t^d + RV_{t-1}^d + \cdots + RV_{t-21}^d)/22 \tag{3}$$

2.2.2. Continuous and Jump Volatility

Based on prior research, this paper decomposes realized volatility into two components—continuous sample path and discrete jump variance—and defines the former part as continuous volatility and the latter as jump volatility. It does this by applying the quadratic variance theory developed by Barndorff-Nielsen and Shephard [4]. Assume that a stock's log price, $p_t = ln(P_t)$ obeys a continuous time jump-diffusion process, which is represented in random difference form as:

$$dp_t = \mu_t dt + \sigma_t dW_t + k_t dq_t, 0 \le t \le T \tag{4}$$

where μ_t is a continuous and locally finite drift function; σ_t is a continuous random fluctuation process; W_t is a standard Wiener process; jump intensity k_t has a normal distribution with mean μ_k and variance σ_k^2; and q_t is a Poisson counting process.

The stock market's log-return volatility for period t in discrete prices is no longer a reliable indicator of the integral volatility, which typically also includes a jump volatility component. Given that the quadratic variance cannot be directly observed, Andersen and Bollerslev demonstrated that the realized volatility RV_t^d is a reliable estimate of the quadratic variance QV_t [1].

$$RV_t^d \xrightarrow{N\to\infty} QV_t = [r,r]_t = \int_{t-1}^{t} \sigma_s^2 ds + \sum_{t-1<s\le t} k_s^2 \tag{5}$$

where $\int_{t-1}^{t} \sigma_s^2 ds$ denotes the integral volatility and it is the continuous volatility component of the overall variance of stock market returns; $\sum_{t-1<s\le t} k_s^2$ denotes the jump volatility of the overall variance.

Additionally, Sheppard and Barndorff-Nielsen found that the realized bi-power variation could measure integral volatility [9]. The realized bi-power variation is a reliable approximation of the continuous volatility when N going to infinity.

$$RBV_t = z_1^{-2}\left(\frac{N}{N-2}\right)\sum_{j=3}^{N} |r_{t,j-2}||r_{t,j}| \tag{6}$$

DOI: 10.54254/2754-1169/45/20230305

where $N/(N-2)$is a correction term for the number of intra-day samples; $z_1 = E(Z_t) = \sqrt{\pi/2}$, and Z_t follows a Gaussian distribution. This study applys the Z_t statistic in determining if the realized volatility contains a jump volatility component by referencing the findings of Huang and Tauchen. [10]. The Z_t statistic is expressed as follows:

$$Z_t = \frac{(RV_t^d - RBV_t)/RV_t^d}{\sqrt{(\mu_1^{-4} + 2\mu_1^{-2} - 5)\frac{1}{M}\max\left(1, \frac{RTQ_t}{RBV_t^2}\right)}} \to N(0,1) \tag{7}$$

where $\mu_1 = \sqrt{2/\pi}, RTQ_t$ is realized tri-power variation, and M is the sampling frequency.

Taken together, it can be observed that an estimator of the daily jump volatility of the stock market may be obtained when jumps are considered to be significant:

$$J_t^d = I(Z_t > \emptyset_\alpha)(RV_t^d - RBV_t) \tag{8}$$

where $I(.)$ is a characteristic function, and $I(.)$ takes 1 when the conditions in parentheses are satisfied, and 0, otherwise. α is usually taken as 0.01.

Similarly, expanding the jumps to weekly and monthly, the expressions are:

$$J_t^w = (J_t^d + J_{t-1}^d + \cdots J_{t-4}^d)/5 \tag{9}$$

$$J_t^m = (J_t^d + J_{t-1}^d + \cdots J_{t-21}^d)/22 \tag{10}$$

After gaining jump volatility, daily continuous volatility is the insignificant jump component plus the significant continuous component showed by the following equation:

$$C_t^d = I(Z_t \le \emptyset_\alpha) \cdot RV_t^d + I(Z_t > \emptyset_\alpha) \cdot RBV_t \tag{11}$$

Similarly, weekly continuous volatility and monthly continuous volatility can be expressed as:

$$C_t^w = (C_t^d + C_{t-1}^d + \cdots C_{t-4}^d)/5 \tag{12}$$

$$C_t^m = (C_t^d + C_{t-1}^d + \cdots C_{t-21}^d)/22 \tag{13}$$

2.2.3. Realized Semi-Variance

In addition to taking jumps into account, researchers have looked at the asymmetric effect of volatility and found that positive and negative returns did not appear to have a consistent impact on volatility. Barndorff-Nielsen [6] proposed the concept of realized semi-variance and divided realized volatility into realized positive and negative semi-variance.

$$RSV_t^{d+} = \sum_{i=0}^{N}[r_{t,i}^2 * I(r_{t,i} > 0)] \tag{14}$$

$$RSV_t^{d-} = \sum_{i=0}^{N}[r_{t,i}^2 * I(r_{t,i} < 0)] \tag{15}$$

Similarly, the realized semi-variance can also be considered for weekly and monthly frequency, as follows.

$$RSV_t^{w+} = (RSV_t^{d+} + RSV_{t-1}^{d+} + \cdots + RSV_{t-4}^{d+})/5 \tag{16}$$

DOI: 10.54254/2754-1169/45/20230305

$$RSV_t^{m+} = (RSV_t^{d+} + RSV_{t-1}^{d+} + \cdots + RSV_{t-21}^{d+})/22 \quad (17)$$

$$RSV_t^{w-} = (RSV_t^{d-} + RSV_{t-1}^{d-} + \cdots + RSV_{t-4}^{d-})/5 \quad (18)$$

$$RSV_t^{m-} = (RSV_t^{d-} + RSV_{t-1}^{d-} + \cdots + RSV_{t-21}^{d-})/22 \quad (19)$$

2.2.4. Variation of Signed Jumps

With further research, Patton and Sheppard subtracted realized negative semi-variance from realized positive semi-variance and eliminated the continuous part of it to construct the signed jump variation, SJ_t and similar positive and negative signed jump variations, SJ_t^+, SJ_t^- [7].

$$SJ_t = RSV_t^{d+} - RSV_t^{d-} \quad (20)$$

$$SJ_t^+ = SJ_t * I(RSV_t^{d+} - RSV_t^{d-} > 0) \quad (21)$$

$$SJ_t^- = SJ_t * I(RSV_t^{d+} - RSV_t^{d-} < 0) \quad (22)$$

2.3. Modelling

HAR models have many advantages compared to other models in terms of representing long-term memory and asymmetry. In order to describe investor behavior of different period, the model adds realized volatility on a daily, weekly, and monthly basis. It also introduces the concept of jump into distinguish anomalous volatility. In portraying asymmetry, three main perspectives are analyzed. First, the realized semi-variance is used to divide the daily realized volatility into two components to test the fluctuation's effects brought on by negative and positive returns on the realized volatility. The realized positive semi-variance represents the part of the day with positive log returns, and the realized negative semi-variance represents the part of the day with negative log returns. Finally, it is separated into positive and negative signed jump variation, which are both used to express the overall positive and negative returns of the day and also imply asymmetry. In summary, a total of four models are constructed in this paper, which are shown below:

This paper refers to Corsi's reasearch on the HAR-RV model [3]. This model is essentially an autoregressive model that takes into account how differences in short-, medium-, and long-term investors' investment preferences affect realized volatility.

$$RV_{t+1}^d = \beta_0 + \beta_d RV_t^d + \beta_w RV_t^w + \beta_m RV_t^m + \varepsilon_{t+1} \quad (23)$$

Refering to Andersen's research [4], the HAR-RV-CJ model, which is based on the HAR-RV model, considers the jump component and the continuous part respectively and separates them into abovementioned parts.

$$RV_{t+1}^d = \beta_0 + \gamma_1 C_t^d + \gamma_2 C_t^w + \gamma_3 C_t^m + \gamma_4 J_t^d + \gamma_5 J_t^w + \gamma_6 J_t^m + \varepsilon_{t+1} \quad (24)$$

This paper refers to Barndorff-Nielsen to establish the HAR-RV-RSV model [6]. This model can account for the leverage impact in dynamic volatility and takes into account three different states of realized positive and negative semi-variance.

$$RV_{t+1}^d = \beta_0 + \gamma_1 RSV_t^{d+} + \gamma_2 RSV_t^{d-} + \gamma_3 RSV_t^{w+} + \gamma_4 RSV_t^{w-} + \gamma_5 RSV_t^{m+} + \gamma_6 RSV_t^{m-} + \varepsilon_{t+1} \quad (25)$$

DOI: 10.54254/2754-1169/45/20230305

This paper refers to Patton and Sheppard to develop the HAR-RV-SJPN model [7]. This model takes into account the direction of price fluctuations and the strength of volatility asymmetry.

$$RV_{t+1}^{d} = \beta_0 + \gamma_1 SJ_t^{+} + \gamma_2 SJ_t^{-} + \gamma_3 RBV_t + \beta_w RV_t^{w} + \beta_m RV_t^{m} + \varepsilon_{t+1} \quad (26)$$

2.4. Testing

2.4.1. Volatility Forecasting Methods

For in-sample parameter estimation, it is not appropriate to use simple ordinary least squares to estimate the parameters due to the possible heteroskedasticity of the model residuals. In this paper, OLS with Newey-West is used for parameter estimation of the above models.

In out-of-sample, the more common rolling time window prediction is used in this paper. The method is briefly described as follows: first, the overall data sample needs to be divided into two parts, the estimation samples, as well as the forecast samples, where the estimation samples are used to estimate the model parameters and the forecast samples are used to predict the estimated value of realized volatility on the same day, with overlap between the two parts. For example, if the overall samples are M days, the first m days are divided into estimation samples, and the forecast samples from $m + 1$ days to M days are the forecast samples to be estimated. Then, m is chosen as the fixed length of the estimation samples, and the data from day 1 to day m is substituted into each of the above volatility models as the estimation samples so that the parameter estimates of the model can be obtained, and then the volatility of day $m + 1$ is estimated using the model. Next, the length of the estimation interval is kept constant and the estimation time is shifted back by 1 day. For instance, the above steps are repeated using the data from day 2 to day $m + 1$ as the estimation samples to obtain the volatility forecast for day $m + 2$; and so on until the volatility forecast for day M is obtained. Thus, for each model, it will get $M - m$ sets of volatility forecasts, and the total samples will have the corresponding true values of realized volatility for that day so that comparisons can be done to evaluating how well the models predict the future.

2.4.2. Out-of-Sample Prediction Evaluation Method

A method to test the prediction accuracy of individual models is to use the MCS test. Model confidence sets, or MCS, are smaller sets of models that comprise the top model with a particular level of confidence and it is a model comparison technique established by Hansen et al. [11]. It recognizes the limitations of the data by only eliminating the poorer prediction models, and sometimes it allows more than one model that may be the best. Briefly, a collection of volatility forecasting models consisting of M_0 are subjected to a significance test as part of the MCS test, which is based on the loss function and the models with relatively poor forecasting ability in the set M_0 are removed. Thus, in each test, the null hypothesis is that two volatility forecasting models in M_0 have the same predictive power for future volatility:

$$H_{0,M}: E\left(d_{i,uv,m}\right) = 0, \forall u, v \in M \in M_0 \quad (27)$$

where $d_{i,uv,m}$ denotes the difference of m loss functions i of volatility forecasting models u and v, and these volatility forecasting models are continuously tested using the equivalence test δ_m and the rejection rule eM, and when no model is rejected from the set, the test ends. In this paper, the equivalence statistic T_R is used as the test statistic, and the formula is as follows:

DOI: 10.54254/2754-1169/45/20230305

$$T_R = \max_{u,v \in M} \left(\frac{\bar{d}_{i,uv}}{\sqrt{var(d_{i,uv})}} \right) \tag{28}$$

If the statistic T_R is greater than the given critical value, the null hypothesis is rejected. The p-value indicates that all models have a superior capacity for prediction when it exceeds the common critical value of 0.1; the higher the p-value, the more likely it is that the model has a better capacity for prediction.

3. Results

3.1. Descriptive Statistics

Table 1 demonstrates that all variables exhibit right-bias and high-peak features. The Jarque-Bera statistic is used to determine if the normal distribution hypothesis is upheld, and all of the series notably disagree with this assumption. Under the unit root test, all the time series are smooth and meet the condition of covariance smoothness. The Q statistic is significant at lags 5, 10, and 20, so each time series has obvious autocorrelation, and it can be judged that the realized volatility and other variables have significant long-term memory, so time series analysis can be applied.

Table 1: Descriptive statistics.

Panel A							
Var.	Mean	Median	Max	Min	St.dev.	Skew	Kurt
RV_t^d	1.8832	0.9391	88.9458	0.0486	4.552	11.0337	164.6716
RV_t^w	1.8839	1.0468	56.8085	0.1388	3.6175	8.5596	100.549
RV_t^m	1.8870	1.1363	26.0608	0.2266	2.7682	5.5212	40.8252
RBV_t	1.5127	0.8144	62.8168	0.0466	3.3787	10.2008	137.5468
C_t^d	1.5107	0.8095	62.8168	0.0466	3.3789	10.2009	137.5422
C_t^w	1.5112	0.8565	41.5647	0.1353	2.8177	8.1045	88.8947
C_t^m	1.5132	0.9131	20.896	0.1819	2.2383	5.4297	39.1486
J_t^d	0.3726	0.0714	65.7894	-10.1611	2.2068	19.8012	506.5898
J_t^w	0.3728	0.141	20.341	-2.0523	1.1684	10.0200	125.2270
J_t^m	0.3738	0.1809	5.3504	-0.3457	0.6614	4.6432	28.3953
RSV_t^{d+}	0.8844	0.4386	61.6329	0.0303	2.2318	15.0887	330.1347
RSV_t^{d-}	0.9989	0.4002	81.1998	0.0155	3.1094	14.9130	309.7687
RSV_t^{w+}	0.8851	0.5144	27.9112	0.0737	1.6649	9.5815	123.1330
RSV_t^{w-}	0.9988	0.4837	32.312	0.0505	2.0921	7.8407	84.5712
RSV_t^{m+}	0.8870	0.5629	13.0368	0.1140	1.2908	6.1926	50.2636
RSV_t^{m-}	1.0001	0.5584	13.0524	0.0852	1.5130	4.8669	32.1453
SJ_t^+	0.4249	0.0000	73.4538	0.0000	2.5519	20.3302	514.1173
SJ_t^-	-0.3104	-0.0090	0.0000	-41.1224	1.3427	-20.0895	540.2953

Panel B					
	J-B	Q (5)	Q (10)	Q (20)	ADF
RV_t^d	2695750***	3067.1***	3913.3***	4866.8***	-8.5317***
RV_t^w	993147***	8172.8***	10554.7***	12942.9***	-7.7036***
RV_t^m	157209***	11484.4***	20708.1***	32113.3***	-5.4886***
RBV_t	1875056***	4180.3***	5457.1***	7200.6***	-8.0489***

DOI: 10.54254/2754-1169/45/20230305

Table 1: (Continued).

C_t^d	1874932***	4180.1***	5457.2***	7199.3***	-8.0517***
C_t^w	773613***	8734.9***	11781.4***	15301.4***	-6.8815***
C_t^m	144246***	11592.1***	21175.6***	33563.1***	-5.1715***
J_t^d	25836067***	155.5***	166.7***	177.8***	-11.9917***
J_t^w	1553281***	4208.8***	4343.0***	4428.7***	-10.7312***
J_t^m	74030***	10222.5***	16752.9***	22371.3***	-7.1234***
RSV_t^{d+}	10927688***	2557.6***	3239.5***	3966.8***	-8.2782***
RSV_t^{d-}	9618405***	999.9***	1343.6***	1707.1***	-9.1801***
RSV_t^{w+}	1498415***	8454.1***	11188.5***	13493.4***	-7.4531***
RSV_t^{w-}	698601***	6775.6***	8415.2***	10221.4***	-8.1043***
RSV_t^{m+}	241708***	11474.8***	20615.5***	31291.2***	-6.045***
RSV_t^{m-}	95600***	11313.5***	20320.8***	31579.9***	-5.4775***
SJ_t^+	26618033***	45.0***	65.6***	84.5***	-10.9832***
SJ_t^-	29392934***	394.0***	440.6***	482.1***	-10.6122***

Note: *** denotes significant at 1% confidence level, J-B denotes Jarque-Bera statistic, Q(n) denotes Ljung-Box Q statistic with nth order lag, and ADF denotes unit root test.

3.2. Regression Results

Table 2 shows the in-sample forecasting results of the four models for the CSI 300 ETF. It can be seen that most of the coefficients of the four models are positive, indicating that all of them can fit the volatility curve well. In terms of individual models, the HAR-RV model shows that the daily and weekly realized volatility have stronger explanatory power for the next day's volatility while the monthly realized volatility do not have good predictive power for the next day; HAR-RV-CJ shows the same trend, the daily and weekly components have strong predictive power for the next day and have significant effect. The non-significant jump component of weekly is not very helpful for forecasting; in the HAR-RV-RSV model, unlike the previous ones, the positive and negative realized semi-variance of monthly has the strongest forecasting power for the short term, with significant coefficients. Except for the weekly realized volatility coefficient, which is not significant in the HAR-RV-SJPN model, all coefficients are significant at the 1% confidence level, showing that the majority of the variables in this model have good short-term forecasting ability.

Table 2: Regression results.

	HAR-RV	HAR-RV-CJ	HAR-RV-RSV	HAR-RV-SJPN
β_0	0.33324**	0.26499**	0.36709***	0.24199*
	(3.102)	(2.685)	(3.435)	(2.47)
β_d	0.18941***			
	(6.301)			
β_w	0.60604***			0.61295***
	(13.685)			(15.194)
β_m	0.04288			-0.01786
	(1.058)			(-0.482)
γ_1		0.33816***	0.24226***	-0.28948***
		(8.954)	(6.272)	(-7.385)
γ_2		0.78911***	0.08099	0.86348***
		(12.147)	(1.49)	(12.868)

DOI: 10.54254/2754-1169/45/20230305

Table 2: (Continued).

γ_3		0.25976**	0.14582	0.60366***
		(3.172)	(1.397)	(17.652)
γ_4		-0.28405***	1.19042***	
		(-5.693)	(8.988)	
γ_5		0.01318	0.92333***	
		(0.111)	(3.91)	
γ_6		-1.06923***	-0.99796***	
		(-4.150)	(-3.837)	

Note: ***, **, * denote significance at 1%, 5%, and 10% confidence levels respectively, with t-values in parentheses.

3.3. MCS Test

The significance level of the MCS test was taken as 0.1, and the six groups of loss functions of the four models in the prediction period were evaluated. These parameters were used for the control parameter of the Bootstrap process: Block length is 2 and the number of simulations B equals to 5000. Table 3 displays the test results.

Table 3: MCS test.

	MSE	MAE	QLIKE	R2LOG	HMSE	HMAE
HAR-RV	0.6477	0.3106	0.7441	0.9478	0.9270	0.6993
HAR-RV-CJ	0.7970	**1.0000**	0.5529	0.7422	0.8936	**1.0000**
HAR-RV-RSV	**1.0000**	0.6841	**1.0000**	**1.0000**	**1.0000**	0.7642
HAR-RV-SJPN	0.7477	0.2503	0.1628	0.2436	0.8449	0.5707

Note: The bolded numbers indicate the best predictive power.

According to Table 3, all four models have a good capacity to predict volatility for the CSI 300 ETF because their p-values are all larger than 0.1 for most loss functions. The HAR-RV model and the HAR-RV-SJPN have lower p-values than that of the other two, and there are no instances in which p-values are 1. As a result, these two models are less accurate at forecasting short-term volatility. After a thorough analysis, it can be said that the HAR-RV-CJ model and the HAR-RV-RSV model have excellent short-term volatility forecasting capabilities because their p-values are larger under all six loss functions and they have p-values of 1 in two cases for the former and four cases for the latter, respectively. The HAR-RV-RSV model is the most accurate among the four models and has superior forecasting abilities over the HAR-RV-CJ model.

4. Conclusion

How to increase the accuracy of volatility forecasting has been the subject of research and one of the toughest challenges in financial academia and practice. Volatility forecasting is very significant in many fields. In recent years, volatility forecasting models based on realized volatility have made great progress in this field. Among them, HAR family volatility forecasting models have attracted extensive research in both academic and practical fields by its better economic implications and forecasting effects. Based on this, four HAR family models are used in this paper, and in-sample analysis of the above models is first conducted using the five-minute high-frequency data of the CSI 300 ETF, and the HAR-RV-CJ model and the HAR-RV-RSV model have significantly better forecasting effect than the other two, showing that the changes have improved the forecasting ability. The HAR-RV-RSV model has the best forecasting ability, indicating that the asymmetric effect of volatility plays a particularly crucial part in forecasting.

DOI: 10.54254/2754-1169/45/20230305

Although research in this paper has achieved some results, there are still some shortcomings that need to be improved, and there are still some works that need to be further extended. First, the models constructed in this paper are only applied to the Chinese stock market, and it needs to be further verified whether the models also performs better in other stock markets. Second, even though the models developed have strong predictive power, there is still some error in the prediction of stock market volatility by these models, so there is still a need to further research. In the following step of research, more factors affecting stock market volatility such as macroeconomic factors and investor sentiment will be taken into consideration. Third, in order to make it more relevant to the real financial market, the next stage will be something applicable.

References

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

[2] *Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). Modeling and Forecasting Realized Volatility. Capital Markets: Asset Pricing & Valuation eJournal.*

[3] *Corsi, F. (2008). A Simple Approximate Long-Memory Model of Realized Volatility. Journal of Financial Econometrics, 7, 174-196.*

[4] *Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2007). Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. The Review of Economics and Statistics, 89(4), 701-720.*

[5] *Shephard, N., & Barndorff-Nielsen, O. (2006). Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation. Journal of Financial Econometrics, 4, 1-30.*

[6] *Barndorff-Nielsen, O. E., Kinnebrock, S., & Shephard, N. (2010). Measuring Downside Risk – Realized Semi-variance. In Volatility and Time Series Econometrics: Essays in Honor of Robert Engle (pp. 0). Oxford University Press.*

[7] *Patton, A. J., & Sheppard, K. (2015). Good Volatility, Bad Volatility: Signed Jumps and The Persistence of Volatility. The Review of Economics and Statistics, 97(3), 683-697.*

[8] *Lyocsa, S., & Stasek, D. (2021). Improving stock market volatility forecasts with complete subset linear and quantile HAR models. Expert Systems with Applications, 183, Article 115416.*

[9] *Hong, W. T., & Hwang, E. (2022). Exponentially Weighted Multivariate HAR Model with Applications in the Stock Market. Entropy, 24(7), Article 937.*

[10] *Huang, X., & Tauchen, G. (2005). The Relative Contribution of Jumps to Total Price Variance. Capital Markets: Asset Pricing & Valuation eJournal.*

[11] *Hansen, P. R., Lunde, A., & Nason, J. M. (2010). The Model Confidence Set. Econometrics eJournal.*