Zhang *et al. Financ Innov* (2021) 7:7
https://doi.org/10.1186/s40854-020-00216-y

Financial Innovation

RESEARCH Open Access

Can the Baidu Index predict realized volatility in the Chinese stock market?

Wei Zhang, Kai Yan and Dehua Shen*

*Correspondence:
dhs@tju.edu.cn
College of Management and Economics, Tianjin University, No. 92 Weijin Road, Nankai District, Tianjin 300072, China

Abstract

This paper incorporates the Baidu Index into various heterogeneous autoregressive type time series models and shows that the Baidu Index is a superior predictor of realized volatility in the SSE 50 Index. Furthermore, the predictability of the Baidu Index is found to rise as the forecasting horizon increases. We also find that continuous components enhance predictive power across all horizons, but that increases are only sustained in the short and medium terms, as the long-term impact on volatility is less persistent. Our findings should be expected to influence investors interested in constructing trading strategies based on realized volatility.

Keywords: Realized volatility, HAR model, Baidu Index, Chinese stock market

Introduction

Forecasting return volatility is a crucial task in investment, option pricing, and risk management. There are two main ways of forecasting return volatility: The first employs the implied volatility derived from option prices as a key predictor. Assessing this method, Latané and Rendleman (1976) and Chiras and Manaster (1978) show that implied volatility performs better than historical standard deviations, where implied volatility is based on past volatility—e.g., realized volatility and jumps (Busch et al. 2011; Christensen and Prabhala 1998). The second way is inferring information from historical data and incorporating data into a GARCH-type model (Bollerslev 1986; Glosten et al. 1993) and a stochastic volatility (SV) model (Harvey et al. 1994). However, these model types rely on low-frequency data—i.e., daily, weekly, or monthly data.

Notably, Andersen et al. (2003) found that models using high-frequency data outperform GARCH-type and SV models due to the fact that the low-frequency data omit important intraday information (Carnero et al. 2004). In light of the past frequently of collection of historical data, various variables and models have been proposed. Initially, Andersen and Bollerslev (1998) suggested using the realized volatility (RV), computed by summing the squared intraday returns. Following this, it was accepted that both the GARCH and SV models can be measured at a high-frequency level, but that the RV is more objective (Barndorff-Nielsen and Shephard 2001; Fleming et al. 2003). In any case, the Internet has fundamentally changed information diffusion patterns in the stock market in the time since, such that scholars began to consider the Internet as one of the most important information sources, incorporating this into the prediction models—e.g., Internet news (Chua and Tsiaplias 2018; Zhang et al. 2016), Twitter (Behrendt & Schmidt 2018; Li et al. 2017), Sina Weibo (Jin et al. 2016), Internet stock message boards (Li et al. 2018), and Google Trends (Da et al. 2011; Dimpfl and Jank 2016). Reflecting the present impact of Internet information sources, this paper employs Internet data to forecast stock return volatility.

This paper focuses on the Chinese stock market because this market is dominated by individual investors and there is a large number of "netizens." A recent survey of Shenzhen Stock Exchange (2018) shows that individual investors accounts for 75.1% of the total in Mainland China equities market. By contrast, individual investors account for only 27% and 12.4% of the U.S. equities market (U.S. Securities and Exchange Commission 2013) and the London Stock Exchange (U.K. Office of National Statistics 2020), respectively. According to the 44th China Statistical Report on Internet Development (China Internet Network Information Center 2019), there are about 854 million "netizens" in China. These country-level characteristics provide a rare opportunity to investigate the predictability of individual investors' information-seeking behavior for return volatility, where the Baidu Index is selected as an appropriate proxy for individual investors' information-seeking behavior, given that, as illustrated by Zhang et al. (2013), the Baidu Index provides more authentic, scientific, and objective results than Google Trends.[1] For the empirical design, we consider the constituent stocks of the SSE 50 Index, comparing various forecasting models deriving from Corsi's (2009) heterogeneous autoregressive (HAR) model. The HAR-type models consider multiscaling features of financial data, where different market participant actions generate different volatility components. Thus, HAR-type models not only produce long-memory volatility (over months), but also deliver clear economic interpretations, which perform better than fractional integration models. Notably, standard GARCH and SV models are not able to reproduce all these features.

Specifically, we construct a novel HAR-type model by incorporating the Baidu Index—i.e., the HAR-RSV-B model, which contains positive and negative realized semivariance, to forecast RV. Therefore, our paper contributes to the existing literature in two ways. Firstly, it contributes to the forecasting literature (e.g., Andersen et al. 2007; Corsi and Reno 2009; Shen et al. 2017) by advocating for the use of a novel and superior predictor—i.e., a weighted Baidu Index. In particular, we find that its predictive ability is more accurate in the long-run, which is interesting as most studies analyzing the Internet communication effect focus on the performance of investor attention in the short term (e.g., Audrino et al. 2020; Bollen et al. 2011; Hamid and Heiden 2015; Ramos et al. 2020; Tantaopas et al. 2016; Vozlyublennaia 2014). Secondly, our findings accord with recent studies on the interdependence between Internet-based activities and stock market performance (Ping and Li 2018; Wen et al. 2019; Yuan 2019). Our study analyzes the predictive power of jump and continuous components, semivariance, and signed jumps that coexist with investor attention and provide evidence regarding the mechanisms of continuous (Andersen et al. 2007) and jump components (Martens et al. 2009).

[1] For a detailed illustration, refer to Sect. Realized volatility: Description of Baidu Index of Zhang et al. (2013).

The remainder of this paper is organized as follows. Section Literature review reviews the relevant literature; Sect. Methodology outlines the methodological approach; Sect. Data describes the data used; Sect. Empirical results and discussion presents the results; and Sect. Conclusion concludes.

Literature review

With more and more frequently collected (intraday) historical data becoming commonplace in financial markets, more sophisticated methods of forecasting return volatility have recently been demanded. Although Blair et al. (2001) found that intraday data only provided little added benefit in implied volatility. The empirical results of Martens and Zein (2002) and Pong et al. (2004) indicate that implied volatility is able to forecast at least as accurately as GARCH models using high-frequency data. So, recent studies have identified a trend of convergence between various methods: Koopman et al. (2005) introduced RV into a GARCH model to perfect the forecast performance, Deo et al. (2006) combined an ARFIMA model with a stochastic volatility model to forecast realized volatility, Dobrev and Szerszen (2010) estimated stochastic volatility by realized volatility measures, Hansen et al. (2012) proposed a measurement equation that added the realized measure to the conditional variance of returns, and Shin and Shin (2019) applied a vector error correction model to take advantage of the cointegration relation between realized volatility and implied volatility.

Intraday data contains many forms of disaggregated information that can improve the accuracy of volatility predictions. Andersen et al. (2004) showed that simple time series models based on RV outperform GARCH-class models. In their 2004 study, Barndorff-Nielsen and Shephard produced an asymptotic model to separate quadratic variation into its continuous and jump components. When these two parts are incorporated into the HAR model, the relevant HAR-CJ models appear (Andersen et al. 2007). The literature initially considered jumps to exhibit weak forecasting ability because of their high prevalence and less enduring nature; but continuous components to be exactly opposite (Andersen et al. 2007; Forsberg and Ghysels 2006). However, in finding a small sample bias for bi-power variation in computing jumps, Corsi et al. (2010) proposed that jumps also have a significant impact on future volatility.

Additionally, semivariance is an important measurement. However, since numerous empirical studies (e.g., Chunhachinda et al. 1997; Fama 1965) show that security returns are not symmetrically distributed, a variable is needed to measure the investment risk. Semivariance, as introduced by Markovitz (1959), is one of the common downsides to risk measures (Huang 2008a). However, Choobineh and Branting (1986) specify optimal estimators for semivariance, and semivariance is applied in asset pricing models by Ang et al. (2006) and in portfolio choice by Huang (2008b), as well as in other sectors.

The use of realized volatility has advantages for long-memory models (Koopman et al. 2005). These long-memory fractional integration models were popular in the past (Shin 2018), but, more recently, diverse modifications based on the HAR model have been proposed by the literature. For instance, Corsi and Reno (2009) added negative returns to investigate the asymmetric leverage effect, a number of empirical analyses indicated that leveraged HAR models improve forecasting ability (e.g., Asai et al. 2012; Audrino and Knaus 2016), and Patton and Sheppard (2015) constructed various HAR-type models with realized semivariance and jumps.

Through more recent studies, scholars continued to improve the ability of models to forecast stock market volatility. Wu and Hou (2019) and Yuan (2019) find that time-varying parameters have greater forecasting accuracies than constant parameters, Wang et al. (2019) find that time-varying transition probabilities (TVTPs) also help the Markov-switching heterogeneous autoregressive (MS-HAR) model perform better, Ma et al. (2019) construct a new jump component in the U.S. stock market, and Ping and Li (2018) propose a truncated two-scale realized volatility (TTSRV) estimator as the continuous part of RV.

The study of the determinants of realized volatility is mainly divided into two aspects. The first relates to the investor agent and participant behavior. In this area of study, Lux and Marchesi (1999) found that noise trade can generate large fluctuations in periods of high volatility, Foucault et al. (2011) showed that retail traders contribute to about 23% of volatility in stock returns, and Barber and Odean (2008) discovered that individual investors are net buyers of attention-grabbing stocks. The second aspect is the effects of related factors on volatility. For instance, Peltomäki et al. (2018) estimate three practical innovations of the investor attention variable in equity and currency markets, Andrei and Hasler (2015) find that both attention and uncertainty are key determinants of asset prices, and Hervé et al. (2019) find investor attention and the participant structure of the market to be closely related.

There are two primary methods of measuring investor attention. The first is direct measurement from the asset itself. Avramov et al. (2006) classify informed and uninformed traders by trade sizes. Many attention-grabbing events are proposed and confirmed, like unusual trading volumes and extreme returns (Barber and Odean 2008), and returns and record events of broader market indeces (Yuan 2015; Hu et al. 2020, 2021). The second is indirect proxies related to the asset. As investors now commonly use the Internet as a primary information channel, many recent studies have constructed novel proxies,[2] linking them to investors' psychological biases.

Methodology

This section provides an empirical definition of volatility and of the components extracted from intraday data and the Baidu Index that will be used in our models (i.e., continuous components, semivariance, signed jumps, and investor attention).

Realized volatility

For a given day t and sample frequency $1/N$, the daily realized volatility proposed by Andersen and Bollerslev (1998) is defined as:

$$RV_{t,N} = \sum_{j=2}^{N+1} r_{t,j}^2 \tag{1}$$

[2] Many novel proxies based on Internet information have been described in the Introduction. For brevity, we do not repeat them in this section.

where $r_{t,j} = 100\left(lnP_{t,j} - lnP_{t,j-1}\right)$ is an intraday return $(j = 2, \ldots, N + 1)$ on day t. $P_{t,j}$ is the last price at time j on day t. Therefore, there are N intervals and $N + 1$ intraday closing prices in one trading day. The call market dominates price discovery (Ellul et al. 2009), and is also a part of daily variance. As such, we adjust the realized volatility to:

$$RV_t = RV_{t,N} + r_{t,1}^2 = \sum_{j=1}^{M} r_{t,j}^2 \tag{2}$$

where $r_{t,1} = 100\left(lnP_{t,1} - lnP_{t-1,end}\right)$ is the call auction variance on day t, $P_{t,1}$ is the opening price of continuous trading on day t, and $P_{t-1,end}$ is the closing price on day $t - 1$. RV_t is the daily complete realized volatility on day t. The length of the supplemental return series $r_{t,j}$ is $M = N + 1$.

Jump and continuous components

We employ a standard jump-diffusion process to estimate the log price of the SSE 50 index $p(t)$ on a trading day:

$$dp(t) = \mu(t)dt + \sigma(t)dW_t + \kappa dq_t \tag{3}$$

where $\mu(t)$ and $\sigma(t)$ denote the drift and instantaneous volatility, W_t is a standard Brownian motion and κdq_t is the pure jump component. Barndorff-Nielsen and Shephard (2004) prove that when $M \to \infty$ daily realized volatility is a consistent estimator of quadratic variation QV_t:

$$RV_t \overset{M\to\infty}{\to} QV_t = \int_{t-1}^{t} \sigma_s^2 ds + \sum_{t-1<s\le t} \kappa_s^2 \tag{4}$$

where $\int_{t-1}^{t} \sigma_s^2 ds$ is an integrated variation of the continuous component and $\sum_{t-1<s\le t} \kappa_s^2$ is the jump component. Meanwhile, the continuous component can be estimated by the realized bi-power variation (RBV) proposed by Barndorff-Nielsen and Shephard (2004):

$$RBV_t = \mu_1^{-2} \frac{M}{M-2} \sum_{j=3}^{M} \left|r_{t,j}\right|\left|r_{t,j-2}\right| \tag{5}$$

where $\mu_p = E\left(|Z|_p\right) = 2^{p/2} \frac{\Gamma((p+1)/2)}{\Gamma(1/2)}$ is the mean of the absolute value of a standard normally distributed random variable and RBV is a consistent estimator of integrated variation. Following Barndorff-Nielsen and Shephard (2006) and Huang and Tauchen (2005), we use Z-statistics to test the significance of the jump component:

$$Z_t = \frac{(RV_t - RBV_t)RV_t^{-1}}{\sqrt{(\mu_1^{-4} + 2\mu_1^{-2} - 5)\frac{1}{M}\max(1, \frac{\mathrm{RTQ_t}}{\mathrm{RBV_t^2}})}} \tag{6}$$

where $RTQ_t = M\mu_{4/3}^{-3}(\frac{M}{M-4}) \sum_{j=5}^{M} \left|r_{t,j-4}\right|^{4/3}\left|r_{t,j-2}\right|^{4/3}\left|r_{t,j}\right|^{4/3}$ is the jump-robust realized tri-power quarticity statistic, $\mu_1 = \sqrt{2/\pi}$ and $\mu_{4/3} = 2^{\frac{2}{3}}\Gamma(\frac{7}{6})\Gamma(\frac{1}{2})^{-1}$.

Thus, the jump component J_t can be defined as:

$$J_t = (RV_t - RBV_t) \times I_{[Z_t > \Phi_\alpha]} \tag{7}$$

where $I(\bullet)$ is the indicator function used to identify the significance and the significance threshold α is 0.99, as per Andersen et al. (2007). Thus, the remainder of the realized volatility is continuous variation C_t, which can be calculated as:

$$C_t = RBV_t \times I_{[Z_t > \Phi_\alpha]} + RV_t \times I_{[Z_t \leq \Phi_\alpha]} \tag{8}$$

Semivariance and signed jumps

The realized semivariance is proposed by Barndorff-Nielsen et al. (2008). The negative realized semivariance estimator is defined as:

$$RSV_t^- = \sum_{j=1}^{M} r_{t,j}^2 \times I_{[r_{t,j} < 0]} \tag{9}$$

Whilst the positive realized semivariance estimator is written as:

$$RSV_t^+ = \sum_{j=1}^{M} r_{t,j}^2 \times I_{[r_{t,j} > 0]} \tag{10}$$

The signed jumps defined by Patton (2011) can be constructed as:

$$\Delta J_t = RSV_t^+ - RSV_t^- \tag{11}$$

Furthermore, the signed jumps can be divided into positive signed jumps $\Delta J_t I_{[\Delta J_t > 0]}$ and negative signed jumps $\Delta J_t I_{[\Delta J_t < 0]}$.

Investor attention based on the Baidu Index

The Baidu Index is based on the number of times users search for keywords, such that it reflects the interest of search engine users to content related to keywords. When investors are interested in one stock, they may search for the security name or its company name in a search engine. However, other users, who are not investors, are more likely to search the company name for contact or recruitment information rather than investment information. Therefore, as a proxy for investor attention, the search query volume of a company name is likely to include a lot of noise, such that the Baidu Index of a security name is more effective. Thus, to investigate the attention given to a security market index, we compute the capitalization-weighted sum of the aggregate Baidu Index of market index components, not market index name, as the proxy variable (Zhang and Wang 2015). Because individual investors are more likely to influence the market index fluctuations by dealing stocks than by trading stock index futures, and generally, institutional investors also do not search for stock index futures before trading them. Thus, the proxy variable for investor attention, B_t, is defined as:

$$B_t = \frac{\sum_{c=1}^{S} (cap_{c,t} \bullet \ln(1 + \mathrm{b_{c,t}}))}{\sum_{c=1}^{S} cap_{c,t}} \tag{12}$$

where $cap_{c,t}$ is the market capitalization of component security c in the given market index on day t and $b_{c,t}$ is the Baidu Index of the component security name. S is the number of shares in the market index.

Model specifications

This paper uses 22 models: 11 existing models and 11 models created for this analysis. These new models are nested models, formulated by adding B to previous models.

Model 1: HAR-RV

The HAR model, as proposed by Corsi (2009), forms the basis of all the models used in our research because it reproduces the long-memory effect of asset volatility. It is specified as:

$$RV_{t+1,t+h} = \beta_0 + \beta_1 RV_t + \beta_5 RV_{t-4,t} + \beta_{22} RV_{t-21,t} + \varepsilon_t \tag{13}$$

where h is the forecasting horizon, $RV_{t+1,t+h}$ is the average realized volatility from $t+1$ to $t+h$, and $RV_{t+1,t+h} = (RV_{t+1} + RV_{t+2} + \cdots + RV_{t+h})/h$. The forecasting result considers the last 1-day, 1-week, and 1-month realized variance, which, according to Corsi (2009), correspond to short-term, medium-term and long-term effects.

Model 2: HAR-RV-J

Andersen et al. (2007) developed their HAR-RV-J model to improve forecast accuracy, adding the last daily jump component to the HAR-RV model to produce:

$$RV_{t+1,t+h} = \beta_0 + \beta_1 RV_t + \beta_5 RV_{t-4,t} + \beta_{22} RV_{t-21,t} + \beta_{J1} J_t + \varepsilon_t \tag{14}$$

where J_t is the jump variation on day t, as computed by Eq. (7).

Model 3: HAR-CJ

The HAR-CJ model proposed by Andersen et al. (2007) is also based on the HAR-RV model, disaggregating realized volatility in each horizon into jump and continuous components, as below:

$$RV_{t+1,t+h} = \beta_0 + \beta_{C1} C_t + \beta_{J1} J_t + \beta_{C5} C_{t-4,t} + \beta_{J5} J_{t-4,t} + \beta_{C22} C_{t-21,t} + \beta_{J22} J_{t-21,t} + \varepsilon_t \tag{15}$$

where C_t is the continuous component on day t defined in Eq. (8), $C_{t-4,t}$ is the average continuous variation over the period $[t-4, t]$, and $C_{t-21,t}$ is the average of the month-lag continuous component. $J_{t-4,t}$ and $J_{t-21,t}$ are the average weekly and monthly jumps, respectively.

Model 4: PS

The PS model proposed by Patton and Sheppard (2015) decomposes daily realized volatility into positive and negative realized semivariance, as below:

$$RV_{t+1,t+h} = \beta_0 + \beta_1^- RSV_t^- + \beta_1^+ RSV_t^+ + \beta_5 RV_{t-4,t} + \beta_{22} RV_{t-21,t} + \varepsilon_t \tag{16}$$

where RSV_t^- is the negative realized semivariance defined in Eq. (9) and RSV_t^+ is the positive realized semivariance specified in Eq. (10).

Model 5: PSLev

The PSLev model adds the leverage effect, as defined by Martens et al. (2009) and generated by negative returns, to the PS model. Patton and Sheppard (2015) proposed assessing if the leverage effect leads to a superior significance of the negative realized semivariance. The model is specified as:

$$RV_{t+1,t+h} = \beta_0 + \beta_1^- RSV_t^- + \beta_1^+ RSV_t^+ + \beta_{m1} RV_t I_{[r_t<0]} + \beta_5 RV_{t-4,t} + \beta_{22} RV_{t-21,t} + \varepsilon_t \tag{17}$$

where $RV_t I_{[r_i<0]}$ is the leverage effect and $I_{[r_i<0]}$ is the indicator function that only a negative return is valid for computing realized volatility in Eq. (1).

Model 6: HAR-RSV

The model developed by Patton and Sheppard (2015) divides realized volatility into positive realized semivariance and negative realized semivariance to assess whether positive and negative parts have different impacts on forecasting. The model is specified as:

$$\begin{aligned} RV_{t+1,t+h} = \beta_0 &+ \beta_1^- RSV_t^- + \beta_1^+ RSV_t^+ + \beta_5^- RSV_{t-4,t}^- + \beta_5^+ RSV_{t-4,t}^+ \\ &+ \beta_{22}^- RSV_{t-21,t}^- + \beta_{22}^+ RSV_{t-21,t}^+ + \varepsilon_t \end{aligned} \tag{18}$$

where $RSV_{t-4,t}^+$ and $RSV_{t-4,t}^-$ are average weekly positive and negative semivariance, respectively. $RSV_{t-21,t}^+$ and $RSV_{t-21,t}^-$ are semivariance for the month horizon.

Model 7: HAR-RSV-J

Chen and Ghysels (2011) produce their HAR-RSV-J model by adding the daily lag jump component to the HAR-RSV model, such that this model can be specified as:

$$\begin{aligned} RV_{t+1,t+h} = \beta_0 &+ \beta_1^- RSV_t^- + \beta_1^+ RSV_t^+ + \beta_5^- RSV_{t-4,t}^- + \beta_5^+ RSV_{t-4,t}^+ \\ &+ \beta_{22}^- RSV_{t-21,t}^- + \beta_{22}^+ RSV_{t-21,t}^+ + \beta_{J1} J_t + \varepsilon_t \end{aligned} \tag{19}$$

Model 8: HAR-RV-SJ

The HAR-RV-SJ model investigates the effect of signed jumps by replacing the daily realized volatility with continuous component and signed jumps in HAR-RV models. It is specified as:

$$RV_{t+1,t+h} = \beta_0 + \beta_{\delta J1} \Delta J_t + \beta_{C1} C_t + \beta_5 RV_{t-4,t} + \beta_{22} RV_{t-21,t} + \varepsilon_t \tag{20}$$

where ΔJ_t is the signed jumps on day t, which is defined in Eq. (11).

Model 9: HAR-CSJ

This model is identical to the HAR-CJ model except for the replacement of jump components with signed jumps. We consider longer-period signed jumps than previous HAR-RV-SJ models by specifying that:

$$\begin{aligned} RV_{t+1,t+h} = \beta_0 &+ \beta_{\delta J1} \Delta J_t + \beta_{C1} C_t + \beta_{\delta J5} \Delta J_{t-4,t} + \beta_{C5} C_{t-4,t} \\ &+ \beta_{\delta J22} \Delta J_{t-21,t} + \beta_{C22} C_{t-21,t} + \varepsilon_t \end{aligned} \tag{21}$$

where $\Delta J_{t-4,t}$ and $\Delta J_{t-21,t}$ are week-lag and month-lag signed jumps.

Model 10: HAR-RV-SJd

The HAR-RV-SJd model represents an improvement over the HAR-RV-SJ model by dividing daily signed jumps into positive signed jumps and negative signed jumps, as below:

$$\begin{aligned} RV_{t+1,t+h} = \beta_0 &+ \beta_{\delta J1}^{-} \Delta J_t I_{[\Delta J_t<0]} + \beta_{\delta J1}^{+} \Delta J_t I_{[\Delta J_t>0]} + \beta_{C1} C_t \\ &+ \beta_5 RV_{t-4,t} + \beta_{22} RV_{t-21,t} + \varepsilon_t \end{aligned} \tag{22}$$

where $\Delta J_t I_{[\Delta J_t<0]}$ is the negative daily signed jump and $\Delta J_t I_{[\Delta J_t>0]}$ is the positive daily signed jump.

Model 11: HAR-CSJd

The HAR-CSJd was proposed by Sévi (2014) and considers many previously stated factors, including dividing signed jumps into positive and negative parts, long-period variables and continuous components. It is written as:

$$\begin{aligned} RV_{t+1,t+h} = \beta_0 &+ \beta_{\delta J1}^{-} \Delta J_t I_{[\Delta J_t<0]} + \beta_{\delta J1}^{+} \Delta J_t I_{[\Delta J_t>0]} + \beta_{C1} C_t \\ &\beta_{\delta J5}^{-} \Delta J_{t-4,t} I_{[\Delta J_{t-4,t}<0]} + \beta_{\delta J5}^{+} \Delta J_{t-4,t} I_{[\Delta J_{t-4,t}>0]} + \beta_{C5} C_{t-4,t} \\ &+ \beta_{\delta J22}^{-} \Delta J_{t-21,t} I_{[\Delta J_{t-21,t}<0]} + \beta_{\delta J22}^{+} \Delta J_{t-21,t} I_{[\Delta J_{t-21,t}>0]} + \beta_{C22} C_{t-21,t} + \varepsilon_t \end{aligned} \tag{23}$$

Model 12: HAR-RV-B

The HAR-RV-B model is a new specification that adds investor attention to the HAR-RV model. We concentrate on the forecast accuracy improvement that B provides, by specifying:

$$RV_{t+1,t+h} = \beta_0 + \beta_1 RV_t + \beta_5 RV_{t-4,t} + \beta_{22} RV_{t-21,t} + \beta_B B_t + \varepsilon_t \tag{24}$$

where B_t is the capital-weighted Baidu Index defined in Eq. (12).

Model 13 to 22: $\underline{B}$ Models

We then develop ten further models by adding B_t to Models (2) to (11) to make Models (12) to (22), which all end with "−B." To avoid repetition, we omit the descriptions of these new models, but Table 1 displays the names and IDs of all 22 models.

Model comparison

The model comparison consists of in-sample analysis and out-of-sample analysis, with OLS regression applied to investigate the aptness of a linear explanation. According to Giot and Laurent (2007), an out-of-sample analysis is the only effective way to evaluate forecasting performance in realized volatility. Generally, the DMW statistic, developed by Diebold and Mariano (1995) and West (1996), is widely used within the forecasting literature.

The DMW test needs a loss function to measure the difference between a real value and a forecasted result in the out-of-sample period. As we use a proxy to estimate the volatility

Table 1 Models specifications

Existing models				New models		
ID	Model name	Reference	Equation number	ID	Model name	Equation number
1	HAR-RV	Corsi (2009)	(13)	12	HAR-RV-B	(24)
2	HAR-RV-J	Andersen et al. (2007)	(14)	13	HAR-RV-J-B	
3	HAR-CJ	Andersen et al. (2007)	(15)	14	HAR-CJ-B	
4	PS	Patton and Sheppard (2015)	(16)	15	PS-B	
5	PSLev	Patton and Sheppard (2015)	(17)	16	PSLev-B	
6	HAR-RSV	Patton and Sheppard (2015)	(18)	17	HAR-RSV-B	
7	HAR-RSV-J	Chen and Ghysels (2011)	(19)	18	HAR-RSV-J-B	
8	HAR-RV-SJ	Patton and Sheppard (2015)	(20)	19	HAR-RV-SJ-B	
9	HAR-CSJ	Sévi (2014)	(21)	20	HAR-CSJ-B	
10	HAR-RV-SJd	Patton and Sheppard (2015)	(22)	21	HAR-RV-SJd-B	
11	HAR-CSJd	Sévi (2014)	(23)	22	HAR-CSJd-B	

instead of observing it directly, a robust loss function is needed to rank two competing models unbiasedly (Patton 2011). As a result of its robustness, Patton (2011) proposes the Q-LIKE loss function, which is defined as:

$$L\left(\widehat{\sigma}^2, v\right) = \log\mathrm{v} + \frac{\widehat{\sigma}^2}{\mathrm{v}} \tag{25}$$

where $\widehat{\sigma}^2$ is a conditionally unbiased volatility proxy, such as realized volatility and v is the forecasted volatility. Then, the difference in loss function for Models A and B at time t is defined as:

$$d_{t,\{A,B\}} = L\left(\widehat{\sigma}_t^2, v_t^A\right) - L\left(\widehat{\sigma}_t^2, v_t^B\right) \tag{26}$$

With a given rolling window size, the moving process will compute a series of losses. The DMW statistic is then given by:

$$DM - QLIKE_{\{A,B\}} = \frac{\overline{d}_{\{A,B\}}}{\sqrt{\widehat{Var}\left(\overline{d}_{\{A,B\}}\right)}} \tag{27}$$

where $\overline{d}_{\{A,B\}}$ is the mean of the difference and $\widehat{Var}\left(\overline{d}_{\{A,B\}}\right)$ is an approximate asymptotic standard variance, which can be estimated as:

$$\widehat{Var}\left(\overline{d}\right) = \frac{1}{P}\left(\widehat{\gamma}_0 + 2\sum_{k=1}^{h}\widehat{\gamma}_k\right) \tag{28}$$

where P is the length of the loss series and h is the forecast horizon. γ_k is the autocovariance of d_t, which can be computed by:

$$\widehat{\gamma}_k = \frac{1}{P}\sum_{t=k+1}^{n}\left(d_t - \bar{d}\right)\left(d_{t-k} - \bar{d}\right) \tag{29}$$

However, the DMW statistic is inappropriate when comparing nested models. Clark and West (2007) adjust the mean squared prediction error (MSPE) and propose the CW statistic. The MSPE of a parsimonious model is expected to be smaller than that of a larger model, as an MSPE-adjusted model is needed to account for the noise (Clark and West 2007).

By way of explanation, we take Model B as the larger model which nests the smaller Model A. h-day ahead forecasts are conducted at time t, such that the real value at time $t + h$ is y_{t+h} and the forecasts of the two models are $\widehat{y}_{1t,t+h}$ and $\widehat{y}_{2t,t+h}$ with corresponding forecast errors $y_{t+h} - \widehat{y}_{1t,t+h}$ and $y_{t+h} - \widehat{y}_{2t,t+h}$. Generally, the sample MSPE is computed by $\left(y_{t+h} - \widehat{y}_{1t,t+h}\right)^2$ and $\left(y_{t+h} - \widehat{y}_{2t,t+h}\right)^2$. Improving on this form, the adjusted MSPE is defined as:

$$\widehat{f}_{t+h} = \left(y_{t+h} - \widehat{y}_{1t,t+h}\right)^2 - \left[\left(y_{t+h} - \widehat{y}_{2t,t+h}\right)^2 - \left(\widehat{y}_{1t,t+h} - \widehat{y}_{2t,t+h}\right)^2\right] \tag{30}$$

Letting $\bar{f}$ be the sample average of $\widehat{f}_{t+h}$, the test statistic becomes:

$$\frac{\sqrt{P}\,\bar{f}}{\sqrt{var\left(\widehat{f}_{t+h} - \bar{f}\right)}} \tag{31}$$

where P is the forecasting length. We reject the null hypothesis if the statistic is greater than $+\,1.282$ at a 10% significant level and $+\,1.645$ at a 5% level.

Data

This paper uses data from the SSE 50 Index of China's securities market. The SSE 50 Index contains 50 stocks of the Shanghai Stock Exchange that are sufficiently large in scale and have good liquidity, and are broadly representative of Chinese enterprises. The sampling frequency of realized variance is five minutes because very few frequencies can beat standard five-minute realized volatility measures in forecasting exercises (Liu et al. 2015). We downloaded all five-minute high-frequency price data from the RESSET dataset.

The Baidu Index data is taken from https://index.baidu.com,[3] which supplies separate indices for different client devices and geographical regions. However, we use the complete index from all regions and devices to investigate the attention of the whole market. We downloaded the component security list, security name, and weight on each trading day from the RESSET dataset.

The investor attention B_t, defined in Eq. (12), is a weighted aggregate measure of the Baidu Index for all SSE 50 companies, except those securities whose names are not included in the keyword directory. Figure 1 illustrates the time series of B_t over the

[3] The Baidu Index (https://index.baidu.com/Helper/?tpl=help) provides data updates daily. The previous day's Baidu Index is usually available before the beginning of market trading.

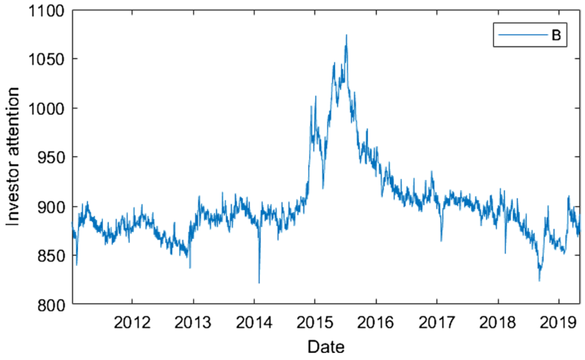

Fig. 1 Investor attention (B) based on Baidu Index

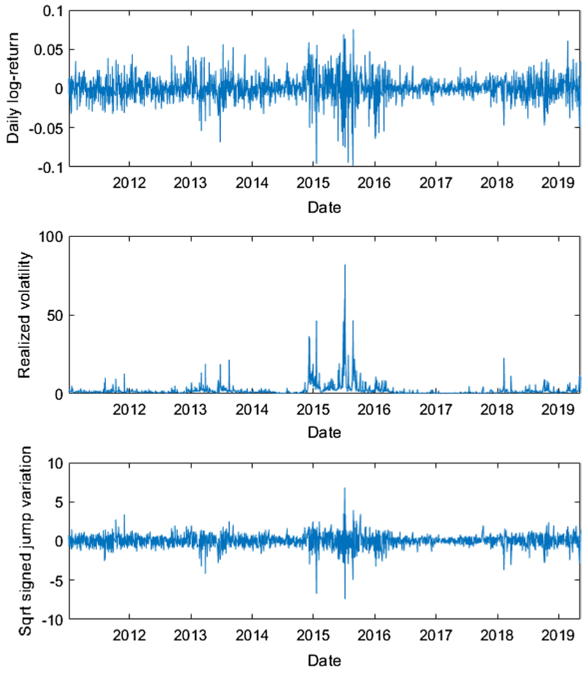

Fig. 2 Log-returns (top panel), realized volatility (middle panel) and sqrt root of signed jump (bottom panel) of SSE 50 over the period. We square the absolute value of signed jumps and keep the sign to reduce the data range

entire sample period. It shows that investor attention boomed in 2015, when the Chinese stock market was experiencing large fluctuations. In other periods, fluctuations are not as exaggerated and occur over shorter periods.

Since Baidu only began publishing its "Baidu Index" product in January 2011, the study's sample period is from January 2011 to May 2019. We remove all non-trading days and obtain 2029 daily observations. Each record contains one realized volatility, RV_t, the jump component, J_t, the continuous component, C_t, the positive semivariance, RSV_t^+, the negative semivariance, RSV_t^-, signed jumps ΔJ_t, and investor attention, B_t, for that day.

The logarithm of daily returns, realized volatility and signed jumps for the SSE 50 are shown by Fig. 2. Periods of relatively low volatility clustering are observed in 2013 and 2018, with a period of very high volatility in 2015. Daily returns and signed jumps are prone to large fluctuations, often moving in unison. Additionally, we find that negative signed jumps are more likely to cause higher volatility in 2013, 2015, and 2018, as would be expected when forecasting short-term volatility.

Figure 3 compares the autocorrelation of realized volatility (RV), positive realized semivariance (RSV^{+}), negative realized semivariance (RSV^{-}), continuous components (C), jump components (J), and signed jumps (ΔJ). The results for RV, RSV^{+}, RSV^{-}, and C all reveal autocorrelation and long-memory processes, but the continuous component displays a more regular autocorrelation. We observe that J and ΔJ are only autocorrelated over only one day, indicating that long-term jumps and signed jumps are almost impossible to predict.

Table 2 reports the statistical properties of all variables for all models. It reveals that the average value of daily, weekly, and monthly variables is approximately equal, but that the variance gradually decreases as the timespan increases. According to the Ljung-Box Q-statistic results, all the variables reject the null hypothesis, and show dynamic dependence at a lag of 5, 10, and 15 days. This phenomenon is beneficial for our regression models. The last column of Table 2 shows the results of an augmented Dickey-Fuller test, which indicates that all variables are stable time series except for monthly mean realized volatility, $RV_{t-21,t}$, the monthly continuous component, $C_{t-21,t}$, and investor attention, B_t.

Empirical results and discussion

This section provides the main results. Firstly, an in-sample analysis of all 22 models forecasting average realized volatility for 1–66 days is provided. We then compare the out-of-sample performance of both existing models and new models. The number of daily observations in our sample is 2008 (from January 2011 to May 2019). These observations are divided into two subgroups: in-sample volatility data covering the first 1000 days and out-of-sample data covering the remaining 1008 days.

In-sample analysis

We estimate Models (1) to (22) introduced in the previous section through OLS regression for h =1–66 (a forecasting horizon ranging from 1 to 66 days that covers the short term, medium term and long term). This provides a clear picture of the performance of each model and the predictive power of various components.

First, Fig. 4 compares the performance of existing models and new models by plotting the mean adjusted R^2 of each model type. These values are high in the short-term but are much lower when the forecast horizon is longer than 15 days. As the time range increases, the gap between existing models and new models is found to widen and the new models perform even better in long-term forecasting. Evidently, it is investor attention (B_t) that improves the precision of forecasts.

Table 3 presents more model-specific results over different time horizons. When predicting the realized volatility of the next day, the results of old models and new models are found to be very similar, as the Baidu Index only improves accuracy in poor models.

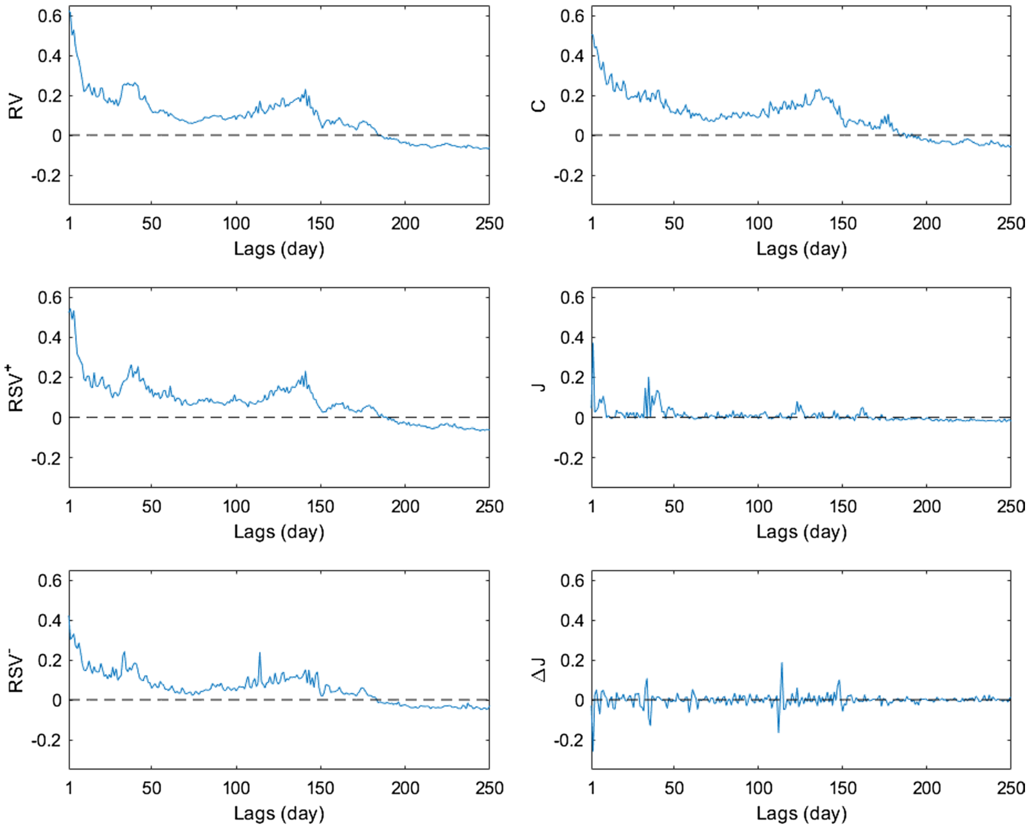

Fig. 3 Sample autocorrelation functions for realized volatility (*RV*), positive realized semivariance (RSV^+), negative realized semivariance (RSV^-), continuous component (*C*), jump component (*J*) and signed jump (ΔJ)

Table 2 Summary statistics for all variables

	Mean	Std	Skewness	Kurtosis	Q(5)	Q(10)	Q(15)	ADF test
RV_t	2.0356	4.3549	8.4772	102.1690	3000.5***	4064.5***	4600.5***	– 19.0158***
$RV_{t-4,t}$	2.0302	3.5615	6.8019	61.1569	7072.3***	9618.7***	10,852***	– 4.9322***
$RV_{t-21,t}$	2.0255	2.8059	4.3552	23.1870	9510.3***	17,147***	22,623***	– 1.5736
J_t	0.3226	2.0344	17.7191	409.9864	103.6***	110.5***	125.0***	– 41.3351***
$J_{t-4,t}$	0.3222	1.0651	9.6433	124.0903	3311.8***	3392.5***	3434.2***	– 11.6409***
$J_{t-21,t}$	0.3199	0.5934	4.3531	24.2468	8381.0***	13,660***	16,480***	– 4.2723***
C_t	1.7130	3.3405	7.0078	65.9576	4002.3***	5655.4***	6548.1***	– 15.6743***
$C_{t-4,t}$	1.7080	2.8584	6.0201	48.5859	7566.3***	10,892***	12,649***	– 4.1946***
$C_{t-21,t}$	1.7056	2.3411	4.2527	24.7815	9598.7***	17,513***	23,343***	– 1.3818
RSV_t^-	1.0354	2.7893	12.2875	225.3263	1110.0***	1633.2***	1890.0***	– 26.0836***
RSV_t^+	1.0002	2.2057	10.5777	183.2157	2551.0***	3307.6***	3640.3***	– 22.1200***
$RSV_{t-4,t}^-$	1.0330	1.9408	6.4176	53.8677	6129.0***	8345.7***	9478.0***	– 6.8134***
$RSV_{t-4,t}^+$	0.9972	1.7340	7.3391	72.0839	7034.2***	9332.2***	10,302***	– 5.2929***
$RSV_{t-21,t}^-$	1.0291	1.4850	4.0589	19.7200	9426.1***	17,057***	22,646***	– 2.0335**
$RSV_{t-21,t}^+$	0.9964	1.3459	4.6139	26.3499	9479.2***	16,941***	22,072***	– 1.6804*
$RV_t I_{[r_i<0]}$	1.0137	3.5338	11.8799	196.9773	486.3***	738.5***	848.2***	– 33.4634***
ΔJ_t	– 0.0352	2.5151	– 5.7525	230.0335	143.8***	163.2***	175.3***	– 46.0637***
$\Delta J_t I_{[\Delta J_t<0]}$	– 0.3867	2.0237	– 17.4996	396.7764	49.81***	85.92***	107.4***	– 39.6309***
$\Delta J_t I_{[\Delta J_t>0]}$	0.3515	1.3994	20.2655	603.1561	229.6***	248.2***	256.0***	– 39.8506***
$\Delta J_{t-4,t}$	– 0.0357	0.9291	– 3.2872	45.5140	1573.0***	1600.9***	1638.4***	– 18.6343***
$\Delta J_{t-4,t} I_{[\Delta J_{t-4,t}<0]}$	– 0.2219	0.7544	– 7.7952	77.4435	2286.1***	2374.6***	2393.1***	– 14.7598***
$\Delta J_{t-4,t} I_{[\Delta J_{t-4,t}>0]}$	0.1861	0.4597	9.0984	145.1430	1075.6***	1107.5***	1113.3***	– 19.9480***
$\Delta J_{t-21,t}$	– 0.0327	0.4005	– 2.2144	8.2741	6750.3***	10,805***	12,539***	– 9.2129***
$\Delta J_{t-21,t} I_{[\Delta J_{t-21,t}<0]}$	– 0.1329	0.3298	– 3.6243	15.3696	6965.5***	11,229***	13,247***	– 8.1011***
$\Delta J_{t-21,t} I_{[\Delta J_{t-21,t}>0]}$	0.1002	0.1581	2.3289	6.6759	6086.9***	9600.1***	11,392***	– 8.6667***
B_t	902.2128	39.2789	1.7131	3.1501	9498.5***	18,584***	27,309***	– 0.1618

Q(5), Q(10) and Q(15) are the Ljung-Box Q-statistics with 5, 10, 15 trading days lag. The last column is augmented Dickey–Fuller test statistic. ***, **, * indicate statistical significance at 1%, 5% or 10% level, respectively

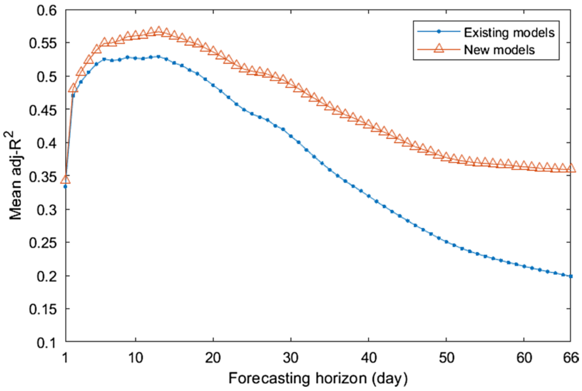

Fig. 4 The mean adjusted R^2 of 11 existing models and 11 new models. Existing models are model (1) to (11) and new models denote model (12) to (22). While calculating the average, the weight of each model is the same

Table 3 The adjusted R^2 of existing models and new models

	Short-term			Medium-term						Long-term								
	h = 1			h = 5			h = 10			h = 22			h = 44			h = 66		
	Before	**After**	**Delta**	**Before**	**After**	**Delta**	**Before**	**After**	**Delta**	**Before**	**After**	**Delta**	**Before**	**After**	**Delta**	**Before**	**After**	**Delta**
HAR-RV-type	0.317	0.331	0.014	0.478	0.509	0.030	0.491	0.535	0.044	0.428	0.497	0.069	0.266	0.394	0.128	0.188	0.357	0.169
HAR-RV-J-type	0.325	0.336	0.011	0.512	0.534	0.022	0.518	0.553	0.035	0.455	0.513	0.058	0.285	0.401	0.116	0.198	0.358	0.160
HAR-CJ-type	0.346	0.354	0.008	0.539	0.556	0.018	0.551	0.580	0.029	0.494	0.543	0.049	0.302	0.412	0.111	0.203	0.361	0.158
PS-type	0.321	0.333	0.012	0.495	0.521	0.026	0.502	0.541	0.039	0.442	0.505	0.063	0.276	0.397	0.121	0.193	0.357	0.164
PSLev-type	0.324	0.335	0.011	0.509	0.533	0.023	0.514	0.551	0.037	0.452	0.512	0.060	0.284	0.401	0.118	0.197	0.359	0.161
HAR-RSV-type	0.339	0.348	0.009	0.519	0.539	0.020	0.531	0.564	0.033	0.477	0.530	0.053	0.291	0.405	0.114	0.198	0.360	0.162
HAR-RSV-J-type	0.340	0.349	0.008	0.531	0.549	0.018	0.542	0.571	0.030	0.485	0.535	0.050	0.297	0.407	0.110	0.201	0.360	0.159
HAR-RV-SJ-type	0.325	0.336	0.011	0.513	0.534	0.022	0.517	0.552	0.035	0.455	0.513	0.057	0.285	0.401	0.115	0.198	0.358	0.160
HAR-CSJ-type	0.347	0.355	0.008	0.541	0.557	0.017	0.553	0.580	0.027	0.497	0.544	0.047	0.303	0.412	0.109	0.205	0.362	0.157
HAR-RV-SJd-type	0.326	0.337	0.011	0.512	0.534	0.022	0.517	0.552	0.035	0.455	0.513	0.058	0.285	0.401	0.116	0.197	0.358	0.160
HAR-CSJd-type	0.356	0.361	0.006	0.544	0.559	0.014	0.555	0.580	0.025	0.499	0.543	0.045	0.308	0.412	0.104	0.208	0.361	0.153

This table presents the adjusted R^2 of existing models (1) to (11) in the column "Before", new models (12) to (22) with investor attention in the column "After" and difference in the column "Delta". The horizon of 1, 5, 10, 22, 44, 66 days cover the short-term, medium-term and long-term

In forecasting medium-term volatility, the Baidu Index plays a more important role, such that new models perform better. The HAR-CSJd-type models perform the best, producing the highest adjusted R^2 values either with or without the Baidu Index, but the gap between HAR-CJ-type, HAR-CSJ-type, and HAR-CSJd-type models is reduced. These three model types with continuous components offer improvements on all other models, whilst the positive and negative semivariance in the HAR-RSV-type and HAR-RSV-J-type models also improve forecasting ability. This confirms the positive impact of disaggregating the realized volatility in prediction.

Finally, we choose the time ranges of 22, 44, and 66 days to assess the accuracy of long-term predictions. As the time horizon increases, the contribution of B_t to all existing models is found to rise, which is consistent with the relationship observed in Fig. 4. For the long-term result, we can still discriminate between models with continuous components, but disparities between new models decrease. The difference between the best new model and the worst new model is 0.050 when $h=5$, but this value falls to 0.005 when $h=66$, which indicates the reduced importance of continuous components.

To be able to draw conclusions about the significance of coefficients, we also consider the estimated parameters of new models. Table 4 reports the estimated result for a 1-day horizon and shows that investor attention is statistically significant at the 5% level for all models. In the HAR-RV-B model, the mean realized volatility of the last day and the last week are significantly positive, but $RV_{t-21,t}$ is not. The HAR-CJ-B model leads to a significant increase in explanatory power due to the decomposition of realized volatility. Jumps have a positive impact on the realized volatility in the short term but the coefficients of jumps over the medium and long term are negative, indicating that they offset the impact of short-term jumps, thereby shadowing the conclusions reached by Andersen et al. (2007).

The coefficient β_{J1} in the HAR-RV-J-B and HAR-RSV-J-B models cannot show the real effect of $J_{t-1,t}$ because realized volatility and semivariance also contain jump factors. As is defined in Eqs. (7) and (8), the realized volatility is the sum of jump and continuous components, such that, for example, the sum of β_1 and β_{J1} is the actual coefficient of the daily jump component of the HAR-RV-J-B model. In Table 3, we show that the HAR-CJ-B, HAR-CSJ-B, and HAR-CSJd-B models with continuous components of each horizon possess the most explanatory power. The 1-day realized volatility is more closely related to the past short-term and medium-term continuous components.

In Table 4, Rows 4–7 report the results for models with positive and negative semivariance. Comparing with the HAR-RV-B model, the decomposition by positive and negative semivariances contributes to the fit of the predictive regression. The 1-day-lagged negative semivariance has a positive effect on the realized volatility, in line with the significance of the downside risk identified by Barndorff-Nielsen et al. (2008). However, interestingly, the positive semivariance of the last week causes higher volatility, but this does not exhibit a strong leverage effect. In the HAR-RV-SJ-B, HAR-CSJ-B, HAR-RV-SJd-B, and HAR-CSJd-B models, signed jumps defined by subtracting negative semivariance from positive semivariance can be used to predict volatility. The higher adjusted R^2 of the HAR-CSJ-B and HAR-CSJd-B models fits the result obtained by Patton and Sheppard (2015), who find that the jump size and sign are the gains from realized jumps. The negative sign of coefficient $\beta_{\delta J1}$ matches the leverage effect of negative semivariance

Table 4 Regression parameters of new models for 1-day horizon

	β_0	β_1	β_5	β_{22}	β_{J1}	β_{J5}	β_{J22}	β_{C1}	β_{C5}	β_{C22}	β_1^-	β_1^+	β_5^-	β_5^+	β_{22}^-	β_{22}^+	β_{m1}	β_B	**Adj-R^2**
HAR-RV-B	− 12.363*** (− 5.785)	0.251*** (9.258)	0.500*** (12.329)	− 0.050 (− 1.226)														0.014*** (5.946)	0.331
HAR-RV-J-B	− 10.361*** (-4.822)	0.393*** (10.758)	0.434*** (10.381)	− 0.040 (− 1.005)	− 0.290*** (− 5.739)													0.012*** (4.963)	0.336
HAR-CJ-B	− 5.852*** (− 2.674)				0.232*** (5.947)	− 0.204** (− 2.113)	− 0.510** (− 2.500)	0.258*** (6.718)	0.751*** (12.570)	0.053 (0.752)								0.007*** (2.784)	0.354
PS-B	− 12.812*** (− 6.049)		0.561*** (13.567)	− 0.063 (− 1.577)							0.372*** (11.247)	− 0.002 (− 0.038)						0.015*** (6.221)	0.333
PSLev-B	− 12.931*** (− 6.113)		0.569*** (13.743)	− 0.070* (− 1.743)							0.229*** (3.615)	0.019 (0.397)					0.117*** (2.663)	0.015*** (6.288)	0.335
HAR-RSV-B	− 12.797*** (− 6.106)										0.441*** (12.830)	− 0.135*** (− 2.624)	− 0.008 (− 0.089)	1.272*** (11.482)	0.184 (0.918)	− 0.360 (− 1.639)		0.015*** (6.271)	0.348
HAR-RSV-J-B	− 9.279*** (− 4.484)				− 0.561*** (− 10.424)						0.814*** (16.610)	− 0.070 (− 1.381)	− 0.188** (− 2.041)	1.310*** (12.132)	0.275 (1.405)	− 0.448** (− 2.090)		0.011*** (4.623)	0.349

	β_0	β_5	β_{22}	β_{C1}	β_{C5}	β_{C22}	$\beta_{\delta J1}$	$\beta_{\delta J1}^-$	$\beta_{\delta J1}^+$	$\beta_{\delta J5}$	$\beta_{\delta J5}^-$	$\beta_{\delta J5}^+$	$\beta_{\delta J22}$	$\beta_{\delta J22}^-$	$\beta_{\delta J22}^+$	β_B	**Adj-R^2**
HAR-RV-SJ-B	− 9.883*** (− 4.700)	0.451*** (11.460)	− 0.048 (− 1.222)	0.396*** (11.083)			− 0.265*** (− 9.715)									0.012*** (4.849)	0.336
HAR-CSJ-B	− 6.725*** (− 3.216)			0.271*** (7.258)	0.793*** (14.870)	− 0.111** (− 2.248)	− 0.368*** (− 12.570)			0.497*** (5.680)			− 0.354* (− 1.822)			0.008*** (3.332)	0.355
HAR-RV-SJd-B	− 9.751*** (− 4.640)	0.475*** (11.630)	− 0.054 (− 1.378)	0.409*** (11.303)				− 0.208*** (− 5.503)	− 0.373*** (− 6.593)							0.011*** (4.791)	0.337
HAR-CSJd-B	− 5.235** (− 2.518)			0.280*** (7.348)	0.742*** (13.753)	− 0.141*** (− 2.899)		− 0.297*** (− 7.858)	− 0.552*** (− 9.207)		0.125 (1.100)	1.256*** (6.494)		− 1.093*** (− 4.282)	1.686*** (3.424)	0.006** (2.497)	0.361

This table reports the in-sample analysis results of the next trading day. New models denote model (12) to (22). Estimation is by OLS and t-statistic is shown in parentheses. ***, **, * indicate statistical significance at 1%, 5% or 10% level, respectively

RSV_t^- in the PS-B, PSLev-B, HAR-RSV-B, and HAR-RSV-J-B models. However, 1-day, 1-week, and 1-month signed jumps have different effects on short-term volatility prediction, which corresponds to the findings from the semivariance. Notably, but perhaps as a result of the nature of the asset assessed, this result contrasts with those obtained by Patton and Sheppard (2015) when analyzing oil future markets. In the stock market, a strong volatility appears likely to follow a positive medium-term semivariance. Thus, overall, the 1-day lagged and 1-week lagged variables are found to be the most important factors in short-term forecasting.

Table 5 reports the in-sample regression result when $h=5$. The 1-month lagged realized volatility is not statistically significant, but the continuous and jump components extracted from this variable are significant, which indicates that the volatility follows a jump process. The HAR-CJ-B, HAR-CSJ-B, and HAR-CSJd-B models, with different horizons of continuous composition, are shown to outperform other new models, confirming the findings of Andersen et al. (2007) that almost all of the predictability in return volatility comes from non-jump components. Yet, we also find evidence that the long-term historical jump components or signed jumps are more important in market volatility forecasting. For 1-week horizon forecasting, the explanatory power of monthly realized volatility is not significant. As the main component of realized volatility, the continuous component C_{22} also has little predictive effect, and it is the monthly jump and signed jumps that contribute the most to the explanatory power. The opposite direction of the coefficient β_{C22} in the HAR-CJ-B and HAR-CSJ-B models also indicates that the jump and signed jumps are more dominant than the continuous component. However, as a result of daily and weekly realized volatility, the effect does not appear in short-term and medium-term parameters. For medium-term forecasting, we note that the coefficients of monthly semivariance and signed jumps are all statistically significant, and exhibit a stronger downside risk effect than signed jumps in other horizons. This result demonstrates that China's stock markets have significant "negative effects" in the long period.

Table 6 reports the estimated parameters for the 1-month horizon. In forecasting long-term volatility, the coefficient of investor attention is larger, but those of other variables are reduced. This change confirms that it is investor attention that narrows the gap between different HAR-type models in volatility forecasting. Many of the short- and medium-term lagged factors are not statistically significant, including 1-day lagged jumps and signed jumps, 5-day lagged semivariance, and realized volatility. However, we note that long-term factors still play a key role in prediction. In addition, comparing the PS-B and the HAR-RSV-B models, we observe that the decomposition of medium-term and long-term semivariance produces a result that is consistent with the long-memory features highlighted by Corsi (2009). We find that the daily signed jump component is insignificant at the 10% level and that the adjusted R^2 of the model is similar to that of the HAR-RV-J-B. This indicates that there is no specific gain to be made from considering signed jumps. However, all the continuous components remain significant with a strong explanatory potential in the long term.

Summarizing the results of the in-sample analysis, we find that investor attention can significantly improve prediction accuracy over the long-term horizon. Comparing different forecast horizons, we find that the range of historical data matches the prediction

Table 5 Regression parameters of new models for 1-week horizon

	β_0	β_1	β_5	β_{22}	β_{J1}	β_{J5}	β_{J22}	β_{C1}	β_{C5}	β_{C22}	β_1^-	β_1^+	β_5^-	β_5^+	β_{22}^-	β_{22}^+	β_{m1}	β_B	**Adj-R^2**
HAR-RV-B	− 18.825*** (− 11.417)	0.256*** (12.259)	0.302*** (9.650)	− 0.029 (− 0.940)														0.022*** (11.757)	0.509
HAR-RV-J-B	− 15.457*** (− 9.629)	0.495*** (18.170)	0.191*** (6.102)	− 0.014 (− 0.456)	− 0.489*** (− 12.970)													0.018*** (9.934)	0.534
HAR-CJ-B	− 9.673*** (− 6.179)				0.164*** (5.886)	− 0.567*** (− 8.237)	− 0.693*** (− 4.756)	0.329*** (12.009)	0.566*** (13.260)	0.132*** (2.601)								0.011*** (6.442)	0.556
PS-B	− 18.567*** (− 11.313)		0.267*** (8.334)	− 0.021 (− 0.689)							0.186*** (7.277)	0.403*** (10.730)						0.022*** (11.646)	0.521
PSLev-B	− 18.776*** (− 11.544)		0.281*** (8.829)	− 0.033 (− 1.078)							− 0.070 (− 1.432)	0.441*** (11.692)					0.208*** (6.182)	0.022*** (11.884)	0.533
HAR-RSV-B	− 18.745*** (− 11.430)										0.194*** (7.218)	0.377*** (9.359)	0.104 (1.435)	0.460*** (5.320)	0.391** (2.498)	− 0.480*** (− 2.792)		0.022*** (11.772)	0.539
HAR-RSV-J-B	− 15.449*** (− 9.678)				− 0.518*** (− 12.458)						0.539*** (14.215)	0.438*** (11.186)	− 0.062 (− 0.873)	0.495*** (5.935)	0.475*** (3.145)	− 0.560*** (− 3.384)		0.018*** (9.992)	0.549

	β_0	β_5	β_{22}	β_{C1}	β_{C5}	β_{C22}	$\beta_{\delta J1}$	$\beta_{\delta J1}^-$	$\beta_{\delta J1}^+$	$\beta_{\delta J5}$	$\beta_{\delta J5}^-$	$\beta_{\delta J5}^+$	$\beta_{\delta J22}$	$\beta_{\delta J22}^-$	$\beta_{\delta J22}^+$	β_B	**Adj-R^2**
HAR-RV-SJ-B	− 15.432*** (− 9.619)	0.192*** (6.394)	− 0.014 (− 0.472)	0.495*** (18.205)			− 0.013 (− 0.621)									0.018*** (9.925)	0.534
HAR-CSJ-B	− 12.575*** (− 7.895)			0.348*** (12.271)	0.493*** (12.151)	− 0.081** (− 2.176)	− 0.046** (− 2.070)			0.190*** (2.858)			− 0.491*** (− 3.321)			0.015*** (8.170)	0.557
HAR-RV-SJd-B	− 15.480*** (− 9.645)	0.183*** (5.868)	− 0.012 (− 0.396)	0.490*** (17.766)				− 0.033 (− 1.162)	0.026 (0.606)							0.018*** (9.951)	0.534
HAR-CSJd-B	− 11.415*** (− 7.161)			0.324*** (11.085)	0.508*** (12.281)	− 0.090** (− 2.414)		− 0.071** (− 2.463)	0.030 (0.653)		0.339*** (3.909)	− 0.262* (− 1.773)		− 1.228*** (− 6.284)	1.524*** (4.043)	0.013*** (7.317)	0.559

This table reports the in-sample analysis results of the next trading week. Other comments are the same as Table 4

Table 6 Regression parameters of new models for 1-month horizon

	β_0	β_1	β_5	β_{22}	β_{J1}	β_{J5}	β_{J22}	β_{C1}	β_{C5}	β_{C22}	β_1^-	β_1^+	β_5^-	β_5^+	β_{22}^-	β_{22}^+	β_{m1}	β_B	**Adj-R^2**
HAR-RV-B	− 29.395*** (− 20.990)	0.104*** (5.836)	0.029 (1.077)	0.110*** (4.176)														0.034*** (21.646)	0.497
HAR-RV-J-B	− 27.240*** (− 19.618)	0.255*** (10.829)	− 0.042 (− 1.558)	0.120*** (4.651)	− 0.310*** (− 9.517)													0.032*** (20.242)	0.513
HAR-CJ-B	− 22.132*** (− 16.031)				0.039 (1.588)	− 0.331*** (− 5.466)	− 0.100*** (− 7.808)	0.156*** (6.426)	0.100*** (2.654)	0.418*** (9.390)								0.026*** (16.609)	0.543
PS-B	− 29.331*** (− 20.939)		0.019 (0.708)	0.112*** (4.249)							0.085*** (3.893)	0.142*** (4.447)						0.034*** (21.592)	0.505
PSLev-B	− 29.349*** (− 20.943)		0.020 (0.748)	0.111*** (4.204)							0.065 (1.546)	0.145*** (4.481)					0.017 (0.577)	0.034*** (21.596)	0.512
HAR-RSV-B	− 29.489*** (− 21.051)										0.069*** (3.013)	0.164*** (4.781)	0.070 (1.127)	− 0.051 (− 0.695)	0.368*** (2.748)	− 0.165 (− 1.123)		0.034*** (21.712)	0.530
HAR-RSV-J-B	− 27.271*** (− 19.652)				− 0.350*** (− 9.744)						0.302*** (9.217)	0.206*** (6.095)	− 0.041 (− 0.673)	− 0.030 (− 0.441)	0.422*** (3.223)	− 0.216 (− 1.507)		0.032*** (20.287)	0.535

	β_0	β_5	β_{22}	β_{C1}	β_{C5}	β_{C22}	$\beta_{\delta J1}$	$\beta_{\delta J1}^-$	$\beta_{\delta J1}^+$	$\beta_{\delta J5}$	$\beta_{\delta J5}^-$	$\beta_{\delta J5}^+$	$\beta_{\delta J22}$	$\beta_{\delta J22}^-$	$\beta_{\delta J22}^+$	β_B	**Adj-R^2**
HAR-RV-SJ-B	− 27.330*** (− 19.673)	− 0.061** (− 2.342)	0.123*** (4.798)	0.259*** (11.015)			− 0.020 (− 1.089)									0.032*** (20.299)	0.513
HAR-CSJ-B	− 25.177*** (− 17.806)			0.165*** (6.564)	0.068* (1.890)	0.141*** (4.270)	− 0.009 (− 0.464)			− 0.012 (− 0.199)			− 0.275** (− 2.092)			0.029*** (18.382)	0.544
HAR-RV-SJd-B	− 27.215*** (− 19.614)	− 0.040 (− 1.500)	0.118*** (4.596)	0.270*** (11.337)				0.028 (1.118)	− 0.109*** (− 2.933)							0.032*** (20.243)	0.513
HAR-CSJd-B	− 24.476*** (− 17.256)			0.176*** (6.793)	0.106*** (2.887)	0.135*** (4.092)		0.038 (1.491)	− 0.081** (− 2.003)		0.161** (2.081)	− 0.406*** (− 3.089)		− 0.708*** (− 4.083)	0.759** (2.245)	0.029*** (17.736)	0.543

This table reports the in-sample analysis results of the next trading month. Other comments are the same as Table 4

Table 7 The CW test between existing models and new models

	h = 1	h = 5	h = 10	h = 22	h = 44	h = 66
HAR-RV-B	0.701*** (3.001)	1.750*** (4.657)	3.318*** (6.146)	4.665*** (7.120)	5.451*** (8.493)	7.748*** (9.674)
HAR-RV-J-B	0.460*** (2.550)	1.109*** (4.108)	2.417*** (5.733)	3.887*** (6.927)	4.835*** (8.188)	7.148*** (9.410)
HAR-CJ-B	0.118 (1.219)	0.397*** (2.969)	1.213*** (5.135)	2.400*** (6.373)	3.349*** (7.240)	5.564*** (8.579)
PS-B	0.814*** (3.217)	1.775*** (4.636)	3.439*** (5.818)	4.862*** (6.689)	5.631*** (8.120)	7.960*** (9.339)
PSLev-B	0.840*** (3.240)	1.837*** (4.736)	3.511*** (5.889)	4.903*** (6.657)	5.677*** (8.090)	8.018*** (9.323)
HAR-RSV-B	0.822*** (2.878)	1.858*** (4.378)	3.604*** (5.648)	5.039*** (6.522)	5.776*** (7.919)	8.055*** (9.172)
HAR-RSV-J-B	0.354** (2.087)	1.139*** (4.002)	2.491*** (5.707)	3.989*** (6.763)	4.914*** (7.989)	7.194*** (9.251)
HAR-RV-SJ-B	0.443*** (2.705)	1.125*** (4.311)	2.472*** (6.064)	3.979*** (7.289)	4.925*** (8.479)	7.254*** (9.645)
HAR-CSJ-B	0.157* (1.454)	0.692*** (3.627)	1.848*** (5.651)	3.229*** (6.774)	4.176*** (7.943)	6.469*** (9.222)
HAR-RV-SJd-B	0.423*** (2.554)	1.127*** (4.243)	2.461*** (5.980)	3.944*** (7.219)	4.879*** (8.382)	7.232*** (9.596)
HAR-CSJd-B	0.056 (0.764)	0.568*** (3.758)	1.649*** (5.744)	3.020*** (6.645)	4.157*** (7.557)	6.122*** (8.789)

A positive difference indicates the new model with investor attention performs better than its correspondent existing model. The forecasting horizon *h* (unit: day) covers short-term, medium-term and long-term. T-statistic is shown in parentheses. ***, ** and * denote the statistical significance at 1%, 5% and 10% level

period. For instance, the future long-term realized volatility depends upon historical monthly components, not 1-day lagged and 1-week lagged variables. This result also confirms the advantages of HAR-type models in forecasting long-term volatility. The decomposition of realized volatility advocated by Andersen et al. (2007) is found to have a significant impact on volatility forecasting (especially the continuous component), but signed jumps perform better than jump components in the SSE 50. Specifically, the HAR-CSJd-B model generates the highest adjusted R^2 over the 1-day and 1-week horizons and the HAR-CSJ-B model produces the highest adjusted R^2 over a 1-month horizon.

Out-of-sample analysis

In this section, we analyze the out-of-sample performance of the 11 existing models and the 11 new models. Specifically, we compare the existing models and their corresponding new models to identify the importance of investor attention. We then compare between different new models for short-term, medium-term, and long-term predictions. A rolling window method is employed to estimate the volatility forecasting results of each model, by adding one new day and removing the most distant day in turn. Therefore, the sample used to estimate the models remains fixed at length $w = 1000$ and the forecasts do not overlap. The number of daily out-of-sample observations is $T = 1008$. For each forecast horizon h, each model will re-estimated $P = T - h + 1$ times, and its parameters are time varying with different samples. Following this process, we produce the loss series of each model with length τ, and evaluate their out-of-sample performance.

Table 7 reports the CW test result for the out-of-sample analysis between existing models and new models. Each new model is the nested model of its correspondent existing model—i.e., the HAR-RV-B model is the larger model which nests the smaller HAR-RV model. There are only two non-significant values in Table 7, which are the HAR-CJ-B and HAR-CSJd-B models for the 1-day forecasting horizon, indicating that the investor attention in these two new models is unable to improve the accuracy of short-term prediction. In addition, the HAR-CJ-B and HAR-CSJd-B models outperform other new models in the in-sample analysis, which indicates that the continuous and jump components have strong predictive power. As the forecasting horizon increases, the gap between existing models and new models widens. Investor attention is thus playing an increasingly important role in volatility forecasting, further verifying the conclusion drawn from the previous analysis.

Next, we compare the out-of-sample performance of new models and report the DMW statistics for various horizons in Tables 8 and 9. Table 8 presents the test result for $h=1$, 5, and 10, which covers the short term and medium term. The results indicate that the differences between the new models are greater: In Panel A, the result obtained at Row HAR-RV-B Column HAR-RV-J-B is 4.0807, which indicates that the HAR-RV-J-B model performs better than the HAR-RV-B model when $h=1$. The PS-B, PSLev-B, and HAR-RSV-B models, which only contain realized volatility and semivariance components, were outperformed by most of the other models, including the original HAR-RV-B model. Furthermore, the decomposition of realized volatility into semivariance does not contribute to volatility forecasting. As expected, given the results of the in-sample analysis, the jump and signed jump indeed play a significant role. The HAR-RSV-J-B model, with the help of the 1-day lagged jump component, outperforms the HAR-RSV-B model.

Considering the 1-week horizon in Panel B, we note that the gap between the models increases and the models with semivariance still do not offer improved performance. The HAR-CJ-B and HAR-CSJd-B model outperform most of the other models, especially in the 1-week lagged and 1-month lagged jumps and signed jumps over the 1-week horizon. At the same time, the HAR-RV-J-B, HAR-RV-SJ-B, and HAR-RV-SJd-B models are outperformed by the HAR-CJ-B and HAR-CSJd-B models. The HAR-CSJ-B model also demonstrates prediction accuracy, but not as effectively as the HAR-CSJd-B model, which indicates that dividing the signed jump into positive and negative aspects is an effective approach.

Panel C shows that the HAR-CSJd-B model is still the most appropriate in the two-week forecasting horizon, but the HAR-CJ-B does not perform as well over the 1-day and 1-week forecasting horizon. The worst models are the PS-B, PSLev-B, and HAR-RSV-B models, which underperform against other models in the short and medium forecasting horizon.

Table 9 reports the DMW statistics for 1-month, two-month and three-month forecasts, with results over the long term differing quite significantly to short-term results. Based on the conclusion that investor attention is a strong predictor over the long term, we note that when $h\geq 22$ all new models mainly rely on the Baidu Index, not the components extracted from realized volatility. In Panel A, the best model is the HAR-CSJ-B model, rather than the HAR-CJ-B or HAR-CSJd-B models. The HAR-RV-J-B model is

Table 8 The DMW statistic for new models in forecasting short-term and medium-term realized volatility

	HAR-RV-J-B	HAR-CJ-B	PS-B	PSLev-B	HAR-RSV-B	HAR-RSV-J-B	HAR-RV-SJ-B	HAR-CSJ-B	HAR-RV-SJd-B	HAR-CSJd-B
Panel A: forecast horizon h = 1										
HAR-RV-B	**4.0807**	1.6891	**− 2.0692**	− 1.8273	**− 2.0006**	0.1823	**3.3036**	1.1339	**2.5973**	1.2350
HAR-RV-J-B		0.3662	**− 3.2676**	**− 3.2847**	**− 3.1182**	− 1.3122	− 1.2267	− 0.2848	− 1.3459	− 0.0502
HAR-CJ-B			**− 2.5581**	**− 2.5172**	**− 4.6960**	**− 4.2706**	− 0.6090	**− 2.7970**	− 0.8017	− 1.1591
PS-B				0.4489	− 0.6842	1.2702	**3.0070**	**2.0891**	**2.8185**	**2.1299**
PSLev-B					− 0.7775	1.1876	**3.0441**	**2.0360**	**2.8454**	**2.0777**
HAR-RSV-B						**3.2200**	**2.9214**	**4.3298**	**2.9654**	**4.2688**
HAR-RSV-J-B							1.0324	**3.4305**	1.0111	**2.6947**
HAR-RV-SJ-B								− 0.0190	− 0.6722	0.1924
HAR-CSJ-B									− 0.1227	0.6201
HAR-RV-SJd-B										0.3405
Panel B: forecast horizon h = 5										
HAR-RV-B	**5.1121**	**4.0309**	1.8631	0.2815	**− 2.5433**	**3.0124**	**5.1543**	**5.4313**	**5.1284**	**5.7115**
HAR-RV-J-B		**2.5138**	**− 4.7525**	**− 5.2395**	**− 4.8428**	**− 3.0511**	1.9244	**2.4414**	− 0.0393	**4.4486**
HAR-CJ-B			**− 3.9631**	**− 4.1078**	**− 4.7136**	**− 3.2531**	**− 2.4783**	− 1.4977	**− 2.5192**	1.8971
PS-B				− 1.7914	**− 3.1354**	**2.3000**	**4.7937**	**4.9920**	**4.7913**	**5.7495**
PSLev-B					**− 2.6953**	**3.0145**	**5.2791**	**5.4161**	**5.2741**	**5.7611**
HAR-RSV-B						**4.3329**	**4.8765**	**5.6750**	**4.8499**	**6.2292**
HAR-RSV-J-B							**3.1623**	**4.5779**	**2.9645**	**5.2586**
HAR-RV-SJ-B								**2.3713**	− 0.3996	**4.4116**
HAR-CSJ-B									**− 2.4615**	**3.6827**
HAR-RV-SJd-B										**4.4581**

Table 8 (continued)

	HAR-RV-J-B	HAR-CJ-B	PS-B	PSLev-B	HAR-RSV-B	HAR-RSV-J-B	HAR-RV-SJ-B	HAR-CSJ-B	HAR-RV-SJd-B	HAR-CSJd-B
Panel C: forecast horizon h = 10										
HAR-RV-B	**3.5871**	1.6059	− 0.2265	0.1255	− 1.5756	0.0165	**3.3574**	**2.5526**	**3.3482**	**3.5180**
HAR-RV-J-B		0.3498	**− 3.5934**	**− 3.2313**	**− 3.2289**	**− 3.1719**	**− 2.2486**	− 0.0045	− 0.9359	**2.1956**
HAR-CJ-B			− 1.6133	− 1.6039	**− 2.1194**	− 1.6005	− 0.5234	− 0.3217	− 0.4417	0.4023
PS-B				0.2670	− 1.6182	0.0692	**3.4080**	**2.6632**	**3.3460**	**3.5591**
PSLev-B					− 1.6408	− 0.0357	**2.9941**	**2.3307**	**3.0113**	**3.3022**
HAR-RSV-B						**2.1674**	**3.0821**	**3.6786**	**3.0989**	**3.7385**
HAR-RSV-J-B							**2.8361**	**4.2123**	**2.8552**	**4.1197**
HAR-RV-SJ-B								0.4644	1.1162	**2.5653**
HAR-CSJ-B									− 0.2333	1.9145
HAR-RV-SJd-B										**2.3833**

Short-term covers forecast horizon of one day. Medium-term covers forecast horizon of 5 and 10 days. The new models are model (12) to (22) with investor attention. A positive statistic indicates that the model in the headline performs better than that in the first column. The statistic is a consistent estimate of the asymptotic variance, whose font is bold for significant result

Table 9 The DMW statistic for new models in forecasting long-term realized volatility

	HAR-RV-J-B	HAR-CJ-B	PS-B	PSLev-B	HAR-RSV-B	HAR-RSV-J-B	HAR-RV-SJ-B	HAR-CSJ-B	HAR-RV-SJd-B	HAR-CSJd-B
Panel A: forecast horizon h = 22										
HAR-RV-B	**2.1857**	− 0.1865	1.7812	0.8469	− 0.6011	0.2462	1.8831	**2.3238**	1.7709	0.8547
HAR-RV-J-B		− 0.5460	**− 2.1105**	**− 2.1356**	− 1.0841	− 0.6966	− 0.6879	1.3434	− 0.4938	− 0.0831
HAR-CJ-B			0.2094	0.1932	− 0.0857	0.2265	0.5237	0.8573	0.5274	0.4166
PS-B				− 1.6391	− 0.6417	0.1731	1.8157	**2.2502**	1.6991	0.7883
PSLev-B					− 0.6173	0.2249	1.8325	**2.3059**	1.7241	0.8383
HAR-RSV-B						1.2031	1.0017	**2.7131**	1.0023	1.6149
HAR-RSV-J-B							0.5903	**2.6321**	0.5933	1.3226
HAR-RV-SJ-B								1.3599	0.0076	− 0.0081
HAR-CSJ-B									− 1.3449	− 1.8402
HAR-RV-SJd-B										− 0.0093
Panel B: forecast horizon h = 44										
HAR-RV-B	1.0521	− 0.0461	− 0.1171	− 0.2264	1.0162	0.9896	0.7485	**2.8947**	0.9737	− 1.0319
HAR-RV-J-B		− 0.2469	− 1.4731	− 1.0271	− 0.4088	0.2733	− 1.3725	**2.9459**	− 1.3741	− 1.4960
HAR-CJ-B			0.0389	0.0380	0.1622	0.2890	0.1828	1.0213	0.2281	− 0.4784
PS-B				− 0.0094	0.9870	1.2398	1.1132	**2.9892**	1.3628	− 1.0251
PSLev-B					0.8211	0.9778	0.7339	**2.7863**	0.9579	− 0.9792
HAR-RSV-B						0.5069	0.0633	**3.0352**	0.3155	− 1.3117
HAR-RSV-J-B							− 0.9668	**2.8459**	− 0.4079	− 1.4716
HAR-RV-SJ-B								**2.9098**	0.8686	− 1.3727
HAR-CSJ-B									**− 2.9845**	**− 2.6827**
HAR-RV-SJd-B										− 1.4652

Table 9 (continued)

	HAR-RV-J-B	HAR-CJ-B	PS-B	PSLev-B	HAR-RSV-B	HAR-RSV-J-B	HAR-RV-SJ-B	HAR-CSJ-B	HAR-RV-SJd-B	HAR-CSJd-B
Panel C: forecast horizon h = 66										
HAR-RV-B	1.1689	− 0.6521	1.6637	1.7833	− 1.0486	− 0.7506	1.3646	0.7236	1.0524	0.5457
HAR-RV-J-B		− 0.9007	− 1.0336	− 1.0653	− 1.4658	− 1.2167	− 0.8408	0.0808	− 0.9470	0.1644
HAR-CJ-B			0.6902	0.6922	0.2102	0.3963	0.7489	1.2456	0.8717	1.0596
PS-B				0.4127	− 1.1371	− 0.8364	0.9754	0.6580	0.8948	0.5046
PSLev-B					− 1.1482	− 0.8481	0.7536	0.6327	0.9139	0.4919
HAR-RSV-B						0.8909	1.2188	**2.6445**	1.3828	1.3024
HAR-RSV-J-B							0.9489	**2.4522**	1.1137	1.1750
HAR-RV-SJ-B								0.5216	0.6664	0.4232
HAR-CSJ-B									− 0.1876	0.2394
HAR-RV-SJd-B										0.2272

Long-term covers forecast horizon of 22, 44 and 66 days. Other comments are the same as Table 8 only more effective than the worst two predictors—the PS-B and PSLev-B models. In Panel B, the HAR-CSJ-B model outperforms other models with significant results. However, the DMW statistic that compares between the HAR-CJ-B and HAR-CSJ-B models is not significant. In Panel C, even the HAR-CSJ-B model only outperforms two models and the jump component does not have a significant predictive impact over the long term, unlike the result of the in-sample analysis.

We conclude that these results are caused by two factors. Firstly, the jump component often derives from macroeconomic events, which makes it difficult to predict and a major driver of short-term volatility. Secondly, the coefficient of the jump component may also be susceptible to external conditions. In the in-sample analysis, all observations are used to evaluate the parameters, but in the out-of-sample analysis, the model trained using historical data is unable to accurately forecast if the condition will change in the future. In addition, the HAR-CSJd-B model is also outperformed by the HAR-CSJ-B model in regards to long-term forecasting. The positive and negative signed jumps can provide more information in short-term and medium-term forecasting but they lead to model overfit for the HAR-CSJd-B model when h is increasing.

To summarize, we conclude that investor attention is valuable in forecasting, but that positive and negative semivariance are not. Furthermore, the in-sample performance can be dramatically improved by disaggregating jump and continuous components over the entire forecasting horizon. However, in long-term forecasting, jumps do not contribute more than other factors extracted from realized volatility, whilst the predictive ability of jumps in long-term forecasting is also affected by other conditions in stock market.

Conclusion

This paper investigates the impact of investor attention on forecasting volatility in the Chinese stock market. Specifically, it adds the Baidu Index as a proxy for investor attention to existing HAR-type models to forecast SSE 50 Index volatility. Using five-minute high-frequency data and collating the Baidu indices of the component security names in the SSE 50 Index, we propose 11 new models by adding the investor attention variable to 11 previously existing models. We then compare their in-sample and out-of-sample predictive power.

The comparison of the models identifies the predictive ability of the variables when taking investor attention into account. The continuous component is found to play an important role in prediction, while the jump component only significantly improves models in the short- and medium-term. Over the long-term horizon, predictive power is reduced by macroeconomic shocks.

It is also shown that investor attention is a useful indicator in forecasting volatility, especially over the long-term horizon. Thus, for security investors, our findings offer an effective risk management and option pricing tool. Specifically, as more option products can be traded in the future, the weighted Baidu Index of component securities will greatly improve the accuracy of original models in predicting long-term volatility. This result is of particular interest because much of the previous research finds the impacts of search query data to be short lived. Consequently, our article provides a new form of evidence within the investor attention research field. Based on our results, it appears feasible that long-term forecasting ability may be related to a discovered long-memory property (Fan et al. 2017), but we leave the analysis of this potential relationship to future research.

Author contributions

All authors have read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China (71790594, 71701150, and U1811462).

Availability of data and materials

The datasets used are available from the corresponding author upon reasonable request.

Competing interests

The authors declare no competing interests.

Received: 12 March 2020 Accepted: 18 November 2020
Published online: 20 January 2021

References

Andersen TG, Bollerslev T (1998) Answering the skeptics: yes, standard volatility models do provide accurate forecasts. Int Econ Rev 39(4):885–905

Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and forecasting realized volatility. Econometrica 71(2):579–625

Andersen TG, Bollerslev T, Meddahi N (2004) Analytical evaluation of volatility forecasts. Int Econ Rev 45(4):1079–1110

Andersen TG, Bollerslev T, Diebold FX (2007) Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. Rev Econ Stat 89(4):701–720

Andrei D, Hasler M (2015) Investor attention and stock market volatility. Rev Financ Stud 28(1):33–72

Ang A, Chen J, Xing Y (2006) Downside risk. Rev Financ Stud 19(4):1191–1239

Asai M, Mcaleer M, Medeiros MC (2012) Asymmetry and long memory in volatility modeling. J Financ Econom 10(3):495–512

Audrino F, Knaus SD (2016) Lassoing the HAR model: a model selection perspective on realized volatility dynamics. Econom Rev 35:1485–1521

Audrino F, Sigrist F, Ballinari D (2020) The impact of sentiment and attention measures on stock market volatility. Int J Forecast 36(2):334–357

Avramov D, Chordia T, Goyal A (2006) The impact of trades on daily volatility. Rev Financ Stud 19(4):1241–1277

Barber BM, Odean T (2008) All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. Rev Financ Stud 21(2):785–818

Barndorff-Nielsen OE, Shephard N (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. J R Stat Soc Ser B (Stat Methodol) 63(2):167–241

Barndorff-Nielsen OE, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps. J Financ Econom 2(1):1–37

Barndorff-Nielsen OE, Shephard N (2006) Econometrics of testing for jumps in financial economics using bipower variation. J Financ Econom 4(1):1–30

Barndorff-Nielsen OE, Kinnebrock S, Shephard N (2008) Measuring downside risk-realised semivariance. CREATES Research Paper (2008-42)

Behrendt S, Schmidt A (2018) The Twitter myth revisited: intraday investor sentiment, Twitter activity and individual-level stock return volatility. J Bank Finance 96:355–367

Blair BJ, Poon SH, Taylor SJ (2001) Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns. J Econom 105(1):5–26

Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econom 31(3):307–327

Busch T, Christensen BJ, Nielsen MO (2011) The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. J Econom 160(1):48–57

Carnero MA, Pena D, Ruiz E (2004) Persistence and kurtosis in GARCH and stochastic volatility models. J Financ Econom 2(2):319–342

Chen X, Ghysels E (2011) News—good or bad—and its impact on volatility predictions over multiple horizons. Rev Financ Stud 24(1):46–81. https://doi.org/10.1093/rfs/hhq071

China Internet Network Information Center (2019) The 44th China statistical report on internet development. http://www.cac.gov.cn/pdf/20190829/44.pdf. Accessed 18 Nov 2020

Chiras DP, Manaster S (1978) The information content of option prices and a test of market efficiency. J Financ Econ 6(2–3):213–234

Choobineh F, Branting D (1986) A simple approximation for semivariance. Eur J Oper Res 27(3):364–370

Christensen BJ, Prabhala NR (1998) The relation between implied and realized volatility. J Financ Econ 50(2):125–150

Chua CL, Tsiaplias S (2018) Information flows and stock market volatility. J Appl Econom 34(1):129–148

Chunhachinda P, Dandapani K, Hamid S, Prakash AJ (1997) Portfolio selection and skewness: evidence from international stock markets. J Bank Finance 21(2):143–167

Clark TE, West KD (2007) Approximately normal tests for equal predictive accuracy in nested models. J Econom 138(1):291–311

Corsi F (2009) A simple approximate long-memory model of realized volatility. J Financ Econom 7(2):174–196

Corsi F, Reno R (2009) HAR volatility modelling with heterogeneous leverage and jumps. Available at SSRN 1316953

Corsi F, Pirino D, Reno R (2010) Threshold bipower variation and the impact of jumps on volatility forecasting. J Econom 159(2):276–288

Da Z, Engelberg J, Gao P (2011) In search of attention. J Finance 66(5):1461–1499

Deo R, Hurvich C, Lu Y (2006) Forecasting realized volatility using a long-memory stochastic volatility model: estimation, prediction and seasonal adjustment. J Econom 131(1–2):29–58

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 20(1):134–144

Dimpfl T, Jank S (2016) Can internet search queries help to predict stock market volatility? Eur Financ Manag 22(2):171–192

Dobrev D, Szerszen P (2010) The information content of high-frequency data for estimating equity return models and forecasting risk. Soc Sci Res Netw 2010(1005):1–42

Ellul A, Shin HS, Tonks I (2009) Opening and closing the market: evidence from the London stock exchange. J Financ Quant Anal 40(4):779–801

Fama EF (1965) Portfolio analysis in a stable Paretian market. Manag Sci 11(3):404–419

Fan X, Yuan Y, Zhuang X, Jin X (2017) Long memory of abnormal investor attention and the cross-correlations between abnormal investor attention and trading volume, volatility respectively. Phys A 469:323–333

Fleming J, Kirby C, Ostdiek B (2003) The economic value of volatility timing using "realized" volatility. J Financ Econ 67(3):473–509

Forsberg L, Ghysels E (2006) Why do absolute returns predict volatility so well. J Financ Econom 5(1):31–67

Foucault T, Sraer D, Thesmar DJ (2011) Individual investors and volatility. J Finance 66(4):1369–1406

Giot P, Laurent S (2007) The information content of implied volatility in light of the jump/continuous decomposition of realized volatility. J Fut Mark 27(4):337–359

Glosten LR, Jagannathan R, Runkle DE (1993) On the relation between the expected value and the volatility of the nominal excess return on stocks. J Finance 48(5):1779–1801

Hamid A, Heiden M (2015) Forecasting volatility with empirical similarity and Google trends. J Econ Behav Organ 117:62–81

Hansen PR, Huang Z, Shek HH (2012) Realized GARCH: a joint model for returns and realized measures of volatility. J Appl Econom 27(6):877–906

Harvey A, Ruiz E, Shephard N (1994) Multivariate stochastic variance models. Rev Econ Stud 61(2):247–264

Hervé F, Zouaoui M, Belvaux B (2019) Noise traders and smart money: evidence from online searches. Econ Model 83:141–149

Hu Y, Li X, Shen D (2020) Attention allocation and international stock return comovement: evidence from the Bitcoin market. Res Int Bus Finance 54:101286

Hu Y, Li X, Goodell JW, Shen D (2021) Investor attention shocks and stock co-movement: substitution or reinforcement? Int Rev Financ Anal 73:101617

Huang XX (2008a) Mean-semivariance models for fuzzy portfolio selection. J Comput Appl Math 217(1):1–8

Huang XX (2008b) Portfolio selection with a new definition of risk. Eur J Oper Res 186(1):351–357

Huang X, Tauchen G (2005) The relative contribution of jumps to total price variance. J Financ Econom 3(4):456–499

Jin X, Shen D, Zhang W (2016) Has microblogging changed stock market behavior? Evidence from China. Phys A 452:151–156

Koopman SJ, Jungbacker B, Hol E (2005) Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements. J Empir Finance 12(3):445–475

Latané HA, Rendleman RJ (1976) Standard deviations of stock price ratios implied in option prices. J Finance 31(2):369–381

Li X, Shen D, Xue M, Zhang W (2017) Daily happiness and stock returns: the case of Chinese company listed in the United States. Econ Model 64:496–501

Li X, Shen D, Zhang W (2018) Do Chinese internet stock message boards convey firm-specific information? Pac Basin Finance J 49:1–14

Liu LY, Patton AJ, Sheppard K (2015) Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. J Econom 187(1):293–311

Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of a financial market. Nature 397(6719):498–500

Ma F, Wahab MIM, Zhang Y (2019) Forecasting the U.S. stock volatility: an aligned jump index from G7 stock markets. Pac Basin Finance J 54:132–146

Markovitz H (1959) Portfolio selection: efficient diversification of investments. Wiley, Hoboken

Martens M, Zein J (2002) Predicting financial volatility: high-frequency time-series forecasts vis-a-vis implied volatility. J Fut Mark 24(11):1005–1028

Martens M, Van Dijk D, De Pooter M (2009) Forecasting S&P 500 volatility: long memory, level shifts, leverage effects, day-of-the-week seasonality, and macroeconomic announcements. Int J Forecast 25(2):282–303

Patton AJ (2011) Volatility forecast comparison using imperfect volatility proxies. J Econom 160(1):246–256

Patton AJ, Sheppard K (2015) Good volatility, bad volatility: signed jumps and the persistence of volatility. Rev Econ Stat 97(3):683–697

Peltomäki J, Graham M, Hasselgren A (2018) Investor attention to market categories and market volatility: the case of emerging markets. Res Int Bus Finance 44:532–546

Ping Y, Li R (2018) Forecasting realized volatility based on the truncated two-scales realized volatility estimator (TTSRV): evidence from China's stock market. Finance Res Lett 25:222–229

Pong SY, Shackleton MB, Taylor SJ, Xu XZ (2004) Forecasting currency volatility: a comparison of implied volatilities and AR(FI)MA models. J Bank Finance 28(10):2541–2563

Ramos SB, Latoeiro P, Veiga H (2020) Limited attention, salience of information and stock market activity. Econ Model 87:92–108
Sévi B (2014) Forecasting the volatility of crude oil futures using intraday data. Eur J Oper Res 235(3):643–659
Shen D, Zhang Y, Xiong X, Zhang W (2017) Baidu index and predictability of Chinese stock returns. Financ Innov. https://doi.org/10.1186/s40854-017-0053-1
Shenzhen Stock Exchange (2018) Individual Investor Status Survey Report: 2017. http://www.szse.cn/aboutus/trends/news/t20180315_519202.html. Accessed 18 Nov 2020
Shin DW (2018) Forecasting realized volatility: a review. J Korean Stat Soc 47(4):395–404
Shin JW, Shin D (2019) Vector error correction heterogeneous autoregressive forecast model of realized volatility and implied volatility. Commun Stat Simul Comput 48(5):1503–1515
Tantaopas P, Padungsaksawasdi C, Treepongkaruna S (2016) Attention effect via internet search intensity in Asia-Pacific stock markets. Pac Basin Finance J 38:107–124
U.K. Office of National Statistics (2020) Ownership of UK quoted shares: 2018. https://www.ons.gov.uk/economy/investmentspensionsandtrusts/bulletins/ownershipofukquotedshares/2018. Accessed 18 Nov 2020
U.S. Securities and Exchange Commission (2013) Institutional Investors: Power and Responsibility. https://www.sec.gov/news/speech/2013-spch041913laahtm#P18_1663. Accessed 18 Nov 2020
Vozlyublennaia N (2014) Investor attention, index performance, and return predictability. J Bank Finance 41:17–35
Wang XX, Shrestha K, Sun Q (2019) Forecasting realised volatility: a Markov switching approach with time-varying transition probabilities. Account Finance 59:1947–1975
Wen F, Xu L, Ouyang G, Kou G (2019) Retail investor attention and stock price crash risk: Evidence from China. Int Rev Financ Anal 65:101376
West KD (1996) Asymptotic inference about predictive ability. Econom J Econom Soc 64:1067–1084
Wu XY, Hou XM (2019) Forecasting realized variance using asymmetric HAR model with time-varying coefficients. Finance Res Lett 30:89–95
Yuan Y (2015) Market-wide attention, trading, and stock returns. J Financ Econ 116(3):548–564
Yuan P (2019) Forecasting realized volatility dynamically based on adjusted dynamic model averaging (AMDA) approach: evidence from China's stock market. J Account Finance 4(2):44
Zhang B, Wang Y (2015) Limited attention of individual investors and stock performance: evidence from the ChiNext market. Econ Model 50:94–104
Zhang W, Shen D, Zhang Y, Xiong X (2013) Open source information, investor attention, and asset pricing. Econ Model 33:613–619
Zhang Y, Song W, Shen D, Zhang W (2016) Market reaction to internet news: information diffusion and price pressure. Econ Model 56:43–49

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.