MÉMOIRE

Realized Variance Forecasting in the US Stock Market

en vue de l'obtention du grade de

Maîtrise en gestion (M. Sc.)
Programme : Ingénierie financière
HEC Montréal

supervisé par

David Ardia

déposé par

Frédéric Rivard

Avril 2024

Remerciements

Je tiens à remercier chaleureusement mon superviseur, David Ardia, pour son soutien constant durant ce projet de recherche. Sa patience et sa compréhension ont été cruciales pour jongler entre mes études, mon travail et mes projets entrepreneuriaux, et pour surmonter les défis organisationnels de ce mémoire.

Un grand merci également à Keven Bluteau et Clément Aymard pour leur aide essentielle dans la collecte et l'organisation des données au début de mes recherches.

Je suis aussi très reconnaissant envers mes parents, ma famille et mes amis pour leur motivation constante durant ma maîtrise. Un merci tout spécial à ma conjointe, Sabrina, pour son soutien sans faille dans les moments clés, qui a grandement contribué à ma réussite!

À la mémoire de Gaétan Aimola

Résumé

Cette mémoire évalue l'efficacité prévisionnelle du modèle Hétérogène AutoRégressif (HAR) et de ses extensions pour la volatilité réalisée sur les marchés financiers. Plus précisément, nous évaluons le modèle HAR ainsi que ses variantes : le modèle HAR avec sauts (HAR-J), le modèle HAR intégrant la semi-variance réalisée (HAR-RSV) et le modèle HAR avec effets de levier (HAR-LE). Ces modèles sont appliqués à l'indice S&P 500 et à 89 titres individuels. Notre analyse comparative révèle que le modèle HAR-LE montre une performance prévisionnelle supérieure pour la variance réalisée, tant en échantillon qu'hors échantillon, particulièrement au niveau de l'indice S&P 500. Nous observons des résultats mitigés concernant les améliorations de la précision prévisionnelle des modèles HAR-J, HAR-RSV, et HAR-LE par rapport au modèle HAR au niveau des titres individuels. Les résultats soulignent le rôle significatif des modèles intégrant des informations sur la volatilité asymétrique, en particulier ceux qui incorporent des effets de levier, pour améliorer la précision des prévisions principalement au niveau de l'indice S&P 500, mais moins au niveau des entreprises individuelles sur le marché boursier américain.

Abstract

This thesis assesses the forecasting effectiveness of the Heterogeneous AutoRegressive (HAR) model and its extensions for realized volatility in financial markets. Specifically, we evaluate the HAR model along with its variants: the HAR model with jumps (HAR-J), the HAR model incorporating realized semivariance (HAR-RSV), and the HAR model with leverage effects (HAR-LE). These models are applied to the S&P 500 Index and 89 individual stocks. Our comparative analysis reveals that the HAR-LE model demonstrates superior forecasting performance for both in-sample and out-of-sample realized variance, particularly at the S&P Index level. We find mixed results with the forecasting accuracy improvements of the HAR-J, HAR-RSV, and HAR-LE compared to the HAR model at the individual firm level. The findings highlight the significant role of models that integrate asymmetric volatility information, especially those that incorporate leverage effects, in improving forecast accuracy primarily at the S&P 500 Index level but less so at the individual firm level within the US stock market.

Contents

List of Figures

List of Tables

Acronyms

ARCH	Autoregressive Conditional Heteroskedasticity
BV	Bipower Variation
BTS	Business Time Sampling
CRSP	Center for Research in Security Prices
DJIA	Dow Jones Industrial Average
EFC	Experiment Flow Chart
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
HAR	Heterogeneous AutoRegressive
HFD	High-Frequency Data
LE	Leverage Effect
MAE	Mean Absolute Error
MAXZ	Maximum Z-score
MINZ	Minimum Z-score
NYSE	New York Stock Exchange
OLS	Ordinary Least Squares
RSV	Realized Semivariance
RV	Realized Variance
SV	Stochastic Volatility
TAQ	Trade and Quote
WRDS	Wharton Research Data Services

1 Introduction

Volatility modeling is indispensable in the financial industry. It impacts crucial areas such as trading, risk management, banking, and government fiscal policy. It equips investors and traders with robust methodologies for forecasting future financial data volatility, facilitating informed strategic decision-making.

Historically, this field has predominantly used volatility models from the Autoregressive Conditional Heteroskedasticity (ARCH) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) families (Engle, 1982; Bollerslev, 1986) as well as the Stochastic Volatility (SV) framework (Shephard, 1996; Taylor, 2007). However, these traditional models often incorporate assumptions that do not fully capture the intricate dynamics of financial markets, particularly the fat-tailed distribution characteristic of financial time series data (Ampadu et al., 2024).

The past two decades have seen a paradigm shift triggered by the influx of high-frequency data (HFD), exposing traditional models' limitations (Andersen and Bollerslev, 1997, 1998; Corsi, 2009). HFD platforms like Wharton Research Data Services (WRDS) and the NYSE's Trade and Quote database (TAQ) provide detailed trade and quote price data that have prompted ongoing refinements in volatility modeling techniques (Shephard and Sheppard, 2010). These enhancements aim to more accurately account for intraday price movements and address the biases inherent in traditional models.

Volatility modeling is one of the main themes of financial research surrounding the advent, accessibility, and availability of HFD (Hussain et al., 2023). A significant breakthrough in this field is the development of the realized variance measure concept in the late 1990s and early 2000s (Andersen et al., 1999, 2000, 2001; Barndorff-Nielsen and Shephard, 2002; Barndorff-Nielsen and Shephard, 2002; Andersen et al., 2003). Realized variance is computed by summing the squared returns of high-frequency financial data over a specified time period. It serves as a reliable estimate of the latent volatility within that time frame. According to Hansen and Lunde (2012a), realized measures built on HFD have substantially enhanced volatility forecasting by improving understanding of the dynamic properties of volatility, serving as effective predictors in reduced-form models, and aiding in developing new models that produce more accurate forecasts.

Realized measures have advanced research and led to the development of the Heterogeneous AutoRegressive (HAR) model of Realized Volatility by Corsi (2009). This model incorpo-

rates data from realized measures from multiple time horizons and adopts a nonparametric approach, thus avoiding the restrictive distributional assumptions inherent in GARCH and SV family models. The HAR model and its extensions that account for asymmetric volatility patterns in equity markets, such as the leverage effect or jump components, have shown effective performance in volatility forecasting. There is ongoing interest in improving volatility forecasting through the framework of the HAR model, with numerous studies focused on stock indices like the S&P 500 and individual stocks. For example, Andersen et al. (2007), Patton and Sheppard (2015), and Lyócsa and Todorova (2020) explore these aspects in the context of the US stock market. Buncic and Gisler (2017) and other authors investigate international markets such as the Nikkei 225 index in Japan (Maki and Ota, 2021) and the PX index in the Czech stock market (Seda, 2012).

Inspired by previous studies and drawing on the groundwork laid by Corsi (2009), this thesis investigates the efficacy of volatility forecasting with the HAR model using the realized variance estimator and high-frequency price data from the NYSE's TAQ database for the S&P 500 Index and 89 individual component stocks over the trading period from 2000 to 2020. Specifically, we assess the forecasting performance of the one-day ahead realized variance in and out-of-sample across two distinct trading subperiods: 2000-2010 and 2011-2020. Our analysis incorporates the HAR model as a benchmark along with three of its extensions designed to account for three types of volatility asymmetries: asymmetric jumps (HAR-J), realized semivariance (HAR-RSV), and the leverage effect (HAR-LE). Our findings contribute to the literature by showing that the HAR-LE model incorporating leverage effect asymmetry dynamics provides superior forecasting accuracy of one-day ahead realized variance at the S&P 500 Index level. However, we find mixed results with the forecasting accuracy of the HAR-J, HAR-RSV, and HAR-LE models compared to the HAR model's accuracy at the individual firm level.

This thesis is structured as follows: Section 2 provides a review of the literature, Section 3 describes the methodology used for the experiment, Section 4 explores the data set of the experiment, Section 5 discusses the results, and Section 6 concludes with a synthesis of the findings and their implications for future research.

2 Literature Review

Volatility forecasting is crucial in financial markets, helping investors and risk managers make informed decisions. Traditionally, models such as the GARCH and SV models have dominated this field. These models, introduced in seminal work by Engle (1982), Bollerslev (1986), and Taylor (2007), are often calibrated on low-frequency data, such as daily price data, to predict future volatility and capture long-term trends.

However, these traditional models often do not account for the nuances of intraday price movements. This can lead to significant biases in volatility estimates by overlooking critical intraday information, as Andersen and Bollerslev (1998) emphasize. Andersen et al. (2003) observe that, with the rise of high-frequency trading and technological advances in the last two decades, there has been a growing need for models that can leverage intraday data to forecast volatility accurately. According to Corsi (2009), traditional models also struggle with issues such as persistence in financial data volatility and their inability to capture the dynamics of volatility across different time horizons, as well as the fat tails and tail crossover phenomena common in financial return distributions. To address these limitations, Corsi (2009) proposes the Heterogeneous AutoRegressive (HAR) model. This model incorporates different volatility components on multiple time horizons and aims to better replicate the empirical features of financial returns, including long memory and fat tails, in a more tractable and parsimonious manner than traditional GARCH and SV models.

The HAR model distinguishes itself by employing a non-parametric approach, avoiding the distributional assumptions required by model families such as GARCH and SV. This allows the HAR model to effectively capture volatility persistence through the aggregation of heterogeneous components, a principle rooted in the Heterogeneous Market Hypothesis and extensively discussed in the literature (Müller et al., 1993, 1997; Dacorogna et al., 1998; Lynch and Zumbach, 2003; Corsi et al., 2012). The model reflects the complex structure of the market, characterized by varying time horizons among participants, whose heterogeneity arises from differences in risk profiles, institutional constraints, access to information, geographical locations, and other characteristics. Central to the HAR model is an ordinary least squares (OLS) regression built on realized variance measure estimators derived from high-frequency intraday financial return data summed and averaged across multiple time frames, typically daily, weekly and monthly.

The simplicity of the HAR model's estimation process and its forecast performance enhance its appeal. Empirical studies have demonstrated the effectiveness of the HAR model in fore-

casting volatility for time series of equity with better performance than traditional models. Thanasoulas (2019) finds that from 2000 to 2018, the HAR model consistently outperformed the GARCH(1,1) model in forecasting volatility in the AEX index in the Netherlands, and the S&P 500 and Nikkei 225 indices, during both non-crisis and crisis periods (2000-2002 and 2007-2009). Furthermore, Seda (2012) shows that during 2004-2012, including the 2008-2009 financial crisis, the HAR model significantly outperformed the GARCH(1,1) model in the in-sample forecast accuracy of the Czech stock market PX index in all periods tested.

Building on the foundational work of Corsi (2009), several extensions of the HAR model have been developed to improve forecast accuracy and accommodate the asymmetric nature of volatility observed in equity markets. Maki and Ota (2021) categorize the expressions of asymmetry in volatility into three principal forms: the leverage effect, which suggests that past negative returns amplify future volatility; realized semivariance (RSV), which splits realized variance into positive and negative components based on intraday returns; and the presence of asymmetric jumps.

The extensive literature shows improvements in volatility forecasting performance when the HAR model is extended to account for one or more of the three forms of volatility asymmetry that Maki and Ota (2021) describe. For example, Buncic and Gisler (2017) examine the impact of jumps and the leverage effect in 18 international equity markets, finding that while the separation of volatility into jump and continuous components shows mixed results, including the leverage effect significantly improves the accuracy of volatility forecasting, particularly in the forecast horizon of one month. Similarly, Corsi and Renò (2012) show that the LHAR-CJ model, which incorporates heterogeneous leverage effects and jumps, significantly enhances volatility prediction across all tested horizons compared to the HAR and HAR-CJ models using high-frequency data from the S&P 500 futures market spanning nearly 28 years (1982 to 2009). Patton and Sheppard (2015), whose findings indicate that negative semivariance is highly predictive of future volatility, report that models that incorporate this measure can greatly enhance forecast accuracy on the S&P 500 index and individual stocks. Lastly, Andersen et al. (2007) observe that including jumps leads to significant improvements in out-of-sample volatility forecasts in exchange rates, equity index returns, and bond yields. This enhancement is attributed to the model's ability to capture abrupt changes in volatility, often linked to specific macroeconomic news announcements or other market-moving events.

3 Methodology

This section discusses the theoretical foundations and practical implications of realized variance. We provide a detailed explanation of each model in our experiment, including the HAR, HAR-J, HAR-RSV, and HAR-LE models, their mathematical framework, and how they are built on realized variance measures. Lastly, we review the key features of HFD and the most commonly used techniques to minimize the influence of market microstructure noise when estimating realized variance.

3.1 Realized Variance

Realized variance[1] is defined as the sum of the squared intraday returns sampled at very high frequencies. This measure is explored in a series of foundational articles, including works by Andersen et al. (1999), Andersen et al. (2000), Andersen et al. (2001), and Andersen et al. (2003). Barndorff-Nielsen and Shephard (2002) and Barndorff-Nielsen and Shephard (2002) further elaborate the theoretical underpinnings of the realized variance measure and use the realized variance to estimate the quadratic variation of a stochastic process. We summarize below the mathematical foundations for the realized variance, based on the established literature and the course notes by Gauthier (2020).

Stock Price Stochastic Process

Let P_t be the price of a stock and $S_t = \ln P_t$. We assume S_t to follow a standard jump-diffusion process,

$$dS_t = \mu_t dt + \sigma_t dW_t + dJ_t \ , \tag{1}$$

where μ_t is the drift, σ_t is the instantaneous stochastic volatility, strictly positive and square-integrable, W_t is a Brownian motion, and J_t is a jump process.

[1] It is important to understand the difference between "realized volatility" and "realized variance". Although these terms are often used interchangeably in the financial literature, their precise meanings differ. Realized volatility refers to the square root of realized variance. To avoid confusion, we clearly define and specify the correct usage of each term in its respective context.

Quadratic Variation of the Stock Price Stochastic Process

Let $\{X_t\}_{t\geq 0}$ be a stochastic process on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $0 = t_0 < t_1^{(n)} < \ldots < t_n^{(n)} = T$ be a partition of the time interval $[0, T]$ such that

$$\lim_{n\to\infty} \max_{i\in\{1,2,\ldots,n\}} \left| t_i^{(n)} - t_{i-1}^{(n)} \right| = 0 \ . \tag{2}$$

Then the quadratic variation of X on the time interval $[0, T]$ is

$$[X]_T = \lim_{n\to\infty} \sum_{i=1}^{n} \left(X_{t_i^{(n)}} - X_{t_{i-1}^{(n)}} \right)^2 \ . \tag{3}$$

It can be shown that the quadratic variation for the logarithmic price process S_t is

$$[S]_T = QV_T = \int_0^T \sigma_s^2 ds + \sum_{i=1}^{N_T} Y_i^2 \ , \tag{4}$$

where $\int_0^T \sigma_s^2 ds$ is the integrated variance and $\sum_{i=1}^{N_T} Y_i^2$ is the sum of the squared jumps.

Daily Realized Variance Estimator of the Stock Price Stochastic Process

Let an intraday return of the stock price process S_t between t and $t + i\frac{\tau}{n}$ be

$$r_{t+i\frac{\tau}{n}} = S_{t+i\frac{\tau}{n}} - S_{t+(i-1)\frac{\tau}{n}} \ . \tag{5}$$

When $\tau = \frac{1}{252}$ is the length in year of a business day and n is the number of periods per day, the realized variance of the stock price process S_t for that specific trading day is

$$RV_{t,t+\tau} = \sum_{i=1}^{n} r_{t+i\frac{\tau}{n}}^2 \ . \tag{6}$$

Assuming the US stock market opening hours from 9:30 am to 4:00 pm, if the intraday returns are calculated at each 1-minute interval period of a trading day, $n = 390$, if they are computed at each five-minute interval period, $n = 78$, etc. It follows that the realized variance is an estimator for the daily quadratic variation and latent volatility of the stochastic

process:

$$\begin{aligned} QV_{t,t+\tau} &= [S]_{t+\tau} - [S]_t \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \left(S_{t+i\frac{\tau}{n}} - S_{t+(i-1)\frac{\tau}{n}} \right)^2 \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} r^2_{t+i\frac{\tau}{n}} \\ &\cong RV_{t,t+\tau} \ . \end{aligned} \tag{7}$$

And, from Equation 4, we can show that

$$\begin{aligned} QV_{t,t+\tau} &= [S]_{t+\tau} - [S]_t \\ &= \underbrace{\underbrace{\int_t^{t+\tau} \sigma_s^2 ds}_{\text{Integrated variance}} + \underbrace{\sum_{i=N_t}^{N_{t+\tau}} Y_i^2}_{\text{Sum of the squared jumps}}}_{\cong RV_{t,t+\tau}} \ . \end{aligned} \tag{8}$$

Bipower Variation

We can disentangle the integrated variance from the realized variance estimator in Equation 8 to estimate the sum of the squared jumps. This requires the bipower variation estimator, which Barndorff-Nielsen and Shephard (2006) show is a consistent estimator of integrated variance. The formula for the bipower variation estimator is

$$BV_{t,t+\tau} = \frac{\pi}{2} \sum_{i=2}^{n} \left| R_{t+i\frac{\tau}{n}} R_{t+(i-1)\frac{\tau}{n}} \right| \ , \tag{9}$$

and we show how the realized variance estimator and the bipower variation estimator are connected in the following equation

$$QV_{t,t+\tau} = \underbrace{\underbrace{\int_t^{t+\tau} \sigma_s^2 ds}_{\text{Integrated variance}}}_{\cong BV_{t,t+\tau}} + \underbrace{\sum_{i=N_t}^{N_{t+\tau}} Y_i^2}_{\text{Sum of the squared jumps}} \ . \tag{10}$$
$$\cong RV_{t,t+\tau}$$

Subtracting the bipower variation from the realized variance provides an estimate for the sum of squared jumps, enhancing the accuracy of jump adjustments in volatility modeling.

Although, in theory, the result of this subtraction should always be positive, in practice, the maximum between 0 and the subtraction result must be taken as illustrated by the formula

$$\sum_{i=N_t}^{N_{t+\tau}} Y_i^2 \cong \max\left(RV_{t,t+\tau} - BV_{t,t+\tau}, 0\right) \ . \tag{11}$$

3.2 Models

We follow the framework of Maki and Ota (2021) to model and forecast realized variance. The models we use are the Heterogeneous AutoRegressive (HAR) and the following extensions: the HAR-J specification, which includes a jump component; the HAR-RSV specification, where realized variance is decomposed into positive and negative realized semivariance; and the HAR-LE specification, which considers the leverage effect of the past negative returns of an asset price time series. The HAR model is our realized variance forecasting benchmark.

In contrast to Corsi (2009) who uses untransformed realized estimators to model volatility under the HAR framework, we apply a logarithmic transformation to the realized variance estimators as Maki and Ota (2021) do. Clements and Preve (2021) explain that this transformation addresses the distributional properties of realized variance. Unlike the untransformed realized variance, which has been observed to be right-skewed, the distribution of the logarithms of realized volatilities is approximately Gaussian (Andersen et al., 2003; Eriksson et al., 2019). Consequently, the logarithmic transformation of the realized variance is shown to be economically and statistically significant (Taylor, 2017).

In each model specification, the realized variance estimator RV_t is constructed on period t equal to one trading day.

3.2.1 HAR

The HAR model aims to address the complexity and persistence seen in the volatility of financial markets by breaking down the volatility into components over different time horizons: daily, weekly, and monthly. To achieve this, the HAR model considers the realized variance observed yesterday (daily), denoted by RV_{t-1}, over the past five trading days (weekly), denoted by RV_{t-1}^w, and over the past twenty-two trading days (approximately one month), denoted by RV_{t-1}^m. The HAR model is defined in the following equation:

$$\ln RV_t = c + \alpha_1 \ln RV_{t-1} + \alpha_2 \ln RV_{t-1}^w + \alpha_3 \ln RV_{t-1}^m + \epsilon_t \tag{12}$$

where $RV_{t-1}^{w} = \frac{1}{5}\sum_{i=1}^{5} RV_{t-i}$ and $RV_{t-1}^{m} = \frac{1}{22}\sum_{i=1}^{22} RV_{t-i}$.

The model captures the cascade of volatility from long-term components to short-term components, which is critical since long-term volatility can significantly influence short-term market risk and vice versa. We use the realized variance estimators under the HAR model specification instead of the realized volatility estimators used by Corsi (2009).

3.2.2 HAR-J

Andersen et al. (2007) note that the HAR model developed by Corsi (2009) can be extended by incorporating a component that accounts for asset price jumps, denoted by J_t. This specification, denoted as the HAR-J model, may help the model better forecast return volatility by separating the jump (discontinuous) component from the continuous component of volatility, the integrated variance. The HAR-J model is defined in the following equation:

$$\ln RV_t = c + \alpha_1 \ln RV_{t-1} + \alpha_2 \ln RV_{t-1}^{w} + \alpha_3 \ln RV_{t-1}^{m} + \beta_1 \ln (J_{t-1} + 1) + \epsilon_t \tag{13}$$

where the jump component $J_t = \max(RV_t - BV_t, 0)$, as previously described in Equation 11.

3.2.3 HAR-RSV

Barndorff-Nielsen et al. (2008) show a method to decompose the realized variance RV_t into two components: the negative realized semivariance, denoted by $RSV_{t=1}^{-}$, and the positive realized semivariance, denoted by RSV_{t-1}^{+}. The authors show that RV_t is the sum of RSV_{t-1}^{-} and RSV_t^{+}. The negative and positive semivariances are computed as follows:

$$\text{the negative realized semivariance, } RSV_{t-1}^{-} = \sum_{j=1}^{n} r_{t,j}^2 I\left\{r_{t,j} < 0\right\} \ , \tag{14}$$

$$\text{and the positive realized semivariance, } RSV_{t-1}^{+} = \sum_{j=1}^{n} r_{t,j}^2 I\left\{r_{t,j} \geq 0\right\} \ .$$

Patton and Sheppard (2015) decompose the daily realized variance estimator variable into positive and negative realized semivariance within the HAR model specification but keep the rest of the model equation intact. The HAR-RSV model is defined in the following equation:

$$\ln RV_t = c + \alpha_1 \ln RSV_{t-1}^{+} + \alpha_2 \ln RSV_{t-1}^{-} + \alpha_3 \ln RV_{t-1}^{w} + \alpha_4 \ln RV_{t-1}^{m} + \epsilon_t \ . \tag{15}$$

The HAR-RSV model aims to capture the leverage effect in financial markets that is not accounted for in the standard HAR model. The model utilizes signed semivariances to differentiate between the content of the information and the risk implications of intraday positive and negative returns.

3.2.4 HAR-LE

Horpestad et al. (2019) describe the HAR-LE model, which accounts for the daily stock return r_t and introduces asymmetry by using leverage terms that are related to lagged absolute daily returns, denoted by $|r_{t-1}|$, and lagged absolute daily negative returns, denoted by $|r_{t-1}| I\{r_{t-1} < 0\}$. The leverage effect can also be extended to longer time horizons by including weekly and monthly returns, as discussed by Corsi and Renò (2009). The HAR-LE model is defined in the following equation:

$$\begin{aligned} \ln RV_t = c &+ \alpha_1 \ln RV_{t-1} + \alpha_2 \ln RV_{t-1}^w + \alpha_3 \ln RV_{t-1}^m \\ &+ \delta_1 |r_{t-1}| + \delta_2 |r_{t-1}| I\{r_{t-1} < 0\} + \epsilon_t \ . \end{aligned} \tag{16}$$

3.3 High-Frequency Data and Sampling

Considering the impact of market microstructure noise when calculating realized variance and bipower variation estimators is crucial. Under ideal conditions, where prices are observed continuously and without measurement error, realized variance can provide an accurate estimate of volatility, as noted by Merton (1980). However, the presence of market microstructure noise in HFD significantly affects the accuracy of the measurement. Specifically, reducing the time interval for calculating intraday returns (by increasing n as shown in Equations 5 and 6) allows for more detailed data aggregation in the computation of the estimators but also leads to increasingly biased estimators. These biases are exacerbated as the sampling frequency increases (Hansen and Lunde, 2012b), often resulting in inflated volatility estimates.

Djupsjöbacka (2010) provides a comprehensive list of microstructure noises present in HFD time series that affect the estimation of the realized variance. These disturbances include bid-ask bounce, nonsynchronous trading, price discreteness and clustering, and market-making activities. The bid-ask bounce causes observed prices to oscillate between bid and ask values, artificially inflating the variability. Nonsynchronous trading refers to delays in recording prices across different markets or assets, leading to misalignments in price data synchronization. Price discreteness and clustering occur when prices aggregate around specific values,

affecting the variance and autocorrelation of returns. Additionally, market-making activities influence measured variance and autocorrelation due to dual-side trading by market makers.

To mitigate the influence of microstructure noise on realized variance estimators, the financial literature suggests various methods, including filtering, outlier detection and removal, smoothing with kernel algorithms, sampling method modifications (e.g., calendar time vs. business time), and subsampling. For instance, Andersen et al. (2001) filter the five-minute interval returns of high-frequency transaction price data on individual stocks from the Dow Jones Industrial Average (DJIA) from January 2, 1993, until May 29, 1998, using a Moving Average model, MA(1), to adjust for the bid-ask bounce and nonsynchronous trading effects. Brownlees and Gallo (2006) apply outlier detection and removal techniques to NYSE TAQ price time series, calculating local statistical baselines within a moving window for each data point and identifying significant deviations using a Z-score type filter. Furthermore, Zhou (1996), Barndorff-Nielsen et al. (2009), and Hansen and Lunde (2004) employ kernel function algorithms to weight and smooth HFD price time series, improving the quality of volatility estimations.

It is important to note that the n parameter in Equations 5 and 6 presumes that intraday returns used to calculate the realized variance estimator come from evenly spaced trade price transactions, which is not always the case in empirical stock price HFD. Practitioners often need to reconstruct the time series of trade prices before estimating the variance realized. Methods such as calendar time sampling, which collects data at regular calendar intervals (e.g., every 5 minutes), business time sampling, which adjusts intervals according to the volume of transactions or trades, and transaction time sampling, which collects data points based on a predefined number of transactions or trading volume, are used to represent market dynamics better. Oomen (2005) suggests that business and transaction time sampling generally provides a more accurate representation of market dynamics compared to calendar time sampling, particularly in reducing the mean square error of bias-corrected realized variance estimates. Extending sampling intervals beyond five minutes to ten or fifteen minutes can mitigate microstructure noise by averaging the effects of order flows and bid-ask bounces, thus providing a clearer view of price trends and reducing the noise-to-signal ratio. Hansen and Lunde (2012b) confirm that noise considerations can be minimized when intraday returns are sampled at lower frequencies, such as every 20 minutes, in the 30 DJIA stocks.

Another method extensively used in the literature to address the challenges of microstructure noise and to find a compromise between frequent sampling and maximizing data utilization

is the subsampling method described by Zhang et al. (2005). This approach averages realized variance estimates computed over different subsampled grids and can reduce the noise impact compared to single-grid methods such as calendar time or business time sampling. The formula for subsampling for the realized variance subsample estimator is

$$RV_{t,t+\tau}^{\text{Subsampling }(k)} = \frac{1}{k}\sum_{j=0}^{k-1} RV_{t+j\Delta,t+\tau}^{\text{Standard }(k)} \tag{17}$$

and for the bipower variation subsample estimator,

$$BV_{t,t+\tau}^{\text{Subsampling }(k)} = \frac{1}{k}\sum_{j=0}^{k-1} BV_{t+j\Delta,t+\tau}^{\text{Standard }(k)} \quad , \tag{18}$$

where k denotes the time interval between two observations and the number of subgrids. For instance, if $k = 5$, it means calculating five subsample estimators at five-minute intervals on each of the five subgrids and then averaging these estimators into one final estimator. The subgrids would be denoted as subgrid 1, $\{5, 10, 15, \ldots, 390\}$, subgrid 2, $\{6, 11, \ldots, 386\}$, and so on. $RV_{t+j\Delta,t+\tau}^{\text{Standard }(1)}$ would the realized variance estimated on subgrid 1, $RV_{t+j\Delta,t+\tau}^{\text{Standard }(2)}$ on the subgrid 2, and so on. Liu et al. (2012) investigates the effectiveness of various high-frequency volatility estimators in multiple asset classes, comparing nearly 400 different realized measures over 11 years of data. They find that while more sophisticated measures sometimes outperform the simple five-minute realized variance, this simpler measure often competes closely regarding practical applicability and forecast accuracy.

4 Data

This section details our data collection methodology, data processing pipeline, and validation procedure with empirical data. We also illustrate some inherent anomalies of high-frequency data in the TAQ database.

4.1 Data Collection

We work with a data set collected by Clément Aymard, a Ph.D. candidate in finance under the supervision of David Ardia. This data set consists of 21 CSV files. The data set includes intraday trade prices at the one-minute interval of stocks that compose the S&P 500 and are traded between 2000 and 2020. There is one CSV file for each trading year. To identify the S&P 500 component stocks for this period, we use the CRSP and NYSE TAQ databases. From CRSP, we obtain unique permanent identifiers (PERMNO) for each of the S&P 500 component stocks. We then locate the corresponding stock data tables with these identifiers in the TAQ database. Furthermore, we focus exclusively on intraday trade price data. Specifically, we do not work with quote price data of identified S&P 500 component stocks also available on TAQ.

4.2 Data Processing Pipeline

We use Python software and various open-source packages to clean and prepare the data, calibrate the models, and analyze the results. To facilitate experiment replication and enable others to follow along, we have uploaded the source code for all manipulation steps to GitHub. The GitHub repository can be made available for cloning upon request at the email address `frederic.rivard@hec.ca`.

The structure of our Python code is organized into four main files. The files `step1.py`, `step2.py`, and `step3.py` each perform a specific stage of the data cleaning pipeline. The file `data_analysis.py` generates the leading figures and results. In the files `step1.py`, `step2.py`, and `step3.py`, basic functions serve as the first level of abstraction and are then integrated into higher-level functions to form the second level of abstraction.

Using Figma software, we visually represent the data cleaning and preparation process in a flow chart. The experiment flow chart (EFC) can be accessed via the following URL: https://www.figma.com/embed?embed_host=share&url=https%3A%2F%2Fwww.figma.com%2Ffile%2FA9npqUxI2BZGctUlb1YCJU%2FData-Pipeline---Realized-Volatility-Forecasting-in-U.

S.-Stock-Markets%3Ftype%3Dwhiteboard%26node-id%3D0%253A1%26t%3DsH6Ua2IG3JITxqpI-1 and is also displayed in Figures 1, 2, and 3. In the EFC chart, the first level of abstraction in function calls is indicated by red folder icons, while the secondary level of abstraction functions is shown with purple rectangles. By examining the legend in Figma and exploring the EFC, readers can easily follow and understand steps 1 to 3 outlined in the subsections below.

4.2.1 Step 1 - Data Preparation and Reorganization

We start by reorganizing the 21 CSV files collected from WRDS, which we identify as Data Set A in the EFC (see Figure 1). More specifically, we identify all unique S&P 500 common stock tickers traded each year between 2000 and 2020. We segregate the 1-minute interval price trade records for each ticker with the `step1.compute_step_1_1()` function and recombine them into 21 annual CSV files. The set of all these segregated trade price data files is called Data Set B in the EFC. We then use the `step1.compute_step_1_2()` function to merge the segregated trade price data on a ticker-by-ticker basis. This merging process reconstructs a complete time series for each unique ticker, extending from its first recorded trade—if it was added to the S&P 500 after January 1, 2000—to its last—if it was removed before December 30, 2020. We obtain a set of 1,139 comprehensive time series, which we call the Datcalla Set C in the EFC.

We observe differences in the structure of the TAQ database between two timeframes, 2000-2009 and 2010-2020, as shown in Tables 1 and 2. During the 2000-2009 period, we use the "SYMBOL" column to differentiate among the S&P 500 component stock tickers. TAQ price trade data for this time period comprise exclusively common stocks. From 2010 onward, the "SYM_ROOT" column replaces "SYMBOL" for ticker identification. A new column called "SYM_SUFFIX" is introduced and displays attributes for the stock symbol of each trade price record, including the funding round (e.g. ticker.A to ticker.T), preferred share status (e.g., ticker.PR), and other characteristics. We consider only those records where the "SYM_SUFFIX" attribute is null, which indicates common stock trade data for the S&P 500 components. For more detailed information on the various suffix types, please consult the NYSE TAQ documentation at https://www.nyse.com/publicdocs/nyse/data/Daily_TAQ_Client_Spec_v4.0.pdf.

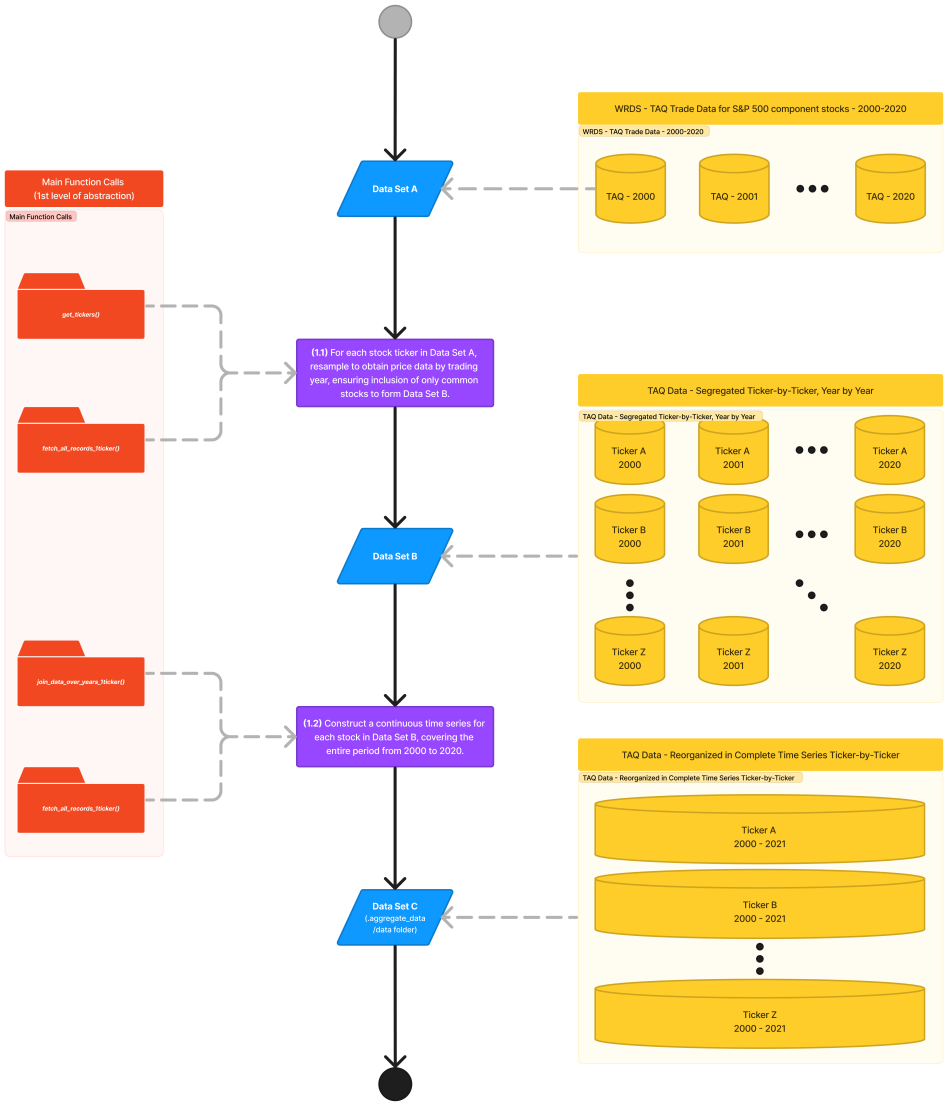

Figure 1: Step 1 - Data Preparation and Reorganization

Table 1: 10 First Rows of TAQ Data Extract: Trading Activity Calendar Year 2000

DATE	SYMBOL	itime	rtime	isize	iprice
20000103	A	9:30:00	9:34:01.0000	64700	78.7500
20000103	A	9:31:00	9:34:01.0000	64700	78.7500
20000103	A	9:32:00	9:34:01.0000	64700	78.7500
20000103	A	9:33:00	9:34:01.0000	64700	78.7500
20000103	A	9:34:00	9:34:01.0000	64700	78.7500
20000103	A	9:35:00	9:34:49.0000	200	78.7500
20000103	A	9:36:00	9:35:58.0000	200	78.6875
20000103	A	9:37:00	9:36:57.0000	1300	78.0625

Table 2: 10 First Rows of TAQ Data Extract: Trading Activity Calendar Year 2010

DATE	SYM_ROOT	SYM_SUFFIX	itime_m	rtime_m	isize	iprice
20100104	A		9:30:00	9:30:02.7640	98	31.32
20100104	A		9:31:00	9:30:51.0490	100	31.14
20100104	A		9:32:00	9:31:59.6660	200	31.36
20100104	A		9:33:00	9:32:54.3190	100	31.23
20100104	A		9:34:00	9:33:59.7630	200	31.22
20100104	A		9:35:00	9:34:51.3600	100	31.28
20100104	A		9:36:00	9:35:58.8740	100	31.30
20100104	A		9:37:00	9:36:57.6640	100	31.33

4.2.2 Step 2 - Data Validation and Cleaning

We validate the integrity and completeness of the raw NYSE TAQ trade price data from 2000 to 2020 in Step 2 of the EFC (see Figure 2). We work only with merged trade price data time series from stocks that trade consistently over this period. To determine which stocks meet this criterion, we use the official trading calendar for the S&P 500 Index, which we collect from a Bloomberg terminal. The trading period between January 1, 2000, and December 30, 2020, consists of 5,282 official trading days. We use the `step2.compute_step_2_1()` function to perform an initial screening of the 1,139 stock time series from Data Set C and find 282 stocks that are traded actively on all official trading days identified.

Aligning stock trading records with Bloomberg's official trading days for S&P 500 Index alone does not suffice to affirm the day-to-day completeness of the TAQ trade price time series. Under normal conditions, a standard trading day from 9:30 am to 4:00 pm should have exactly 390 price observations per stock at a one-minute trading interval. On certain US holidays, such as Independence Day or Thanksgiving, fewer observations are recorded when the NYSE operates on a shortened schedule and closes at 1:00 pm. To ensure each trading day fully captures all trade price observations, we use two additional functions: `step2.compute_step_2_2_1()` and `step2.compute_step_2_2_2()`. These functions verify that there are either 390 observations for a regular full-trading day or ensure that observations are consecutive at one-minute intervals for days with early market closes.

We implement an initial filtering process to eliminate potential errors and ensure the completeness of the TAQ time series daily observations. Our approach follows the guidelines proposed by Hendershott and Moulton (2011), which remove intraday price records from the dataset that do not satisfy certain heuristic criteria. These criteria include trade price records valued at zero or those where the price exceeds 150% or falls below 50% of the previous price. We adapt these principles to fit a five-minute subsample estimation framework, whereby we completely discard any stock ticker time series that contains trade price records meeting these criteria on any trading day and within any five-minute subgrid interval. To facilitate these filtering steps, we use the `step2.compute_step_2_3()` function. We exclude 179 stocks from Data Set C and retain 103 stocks for Data Set D.

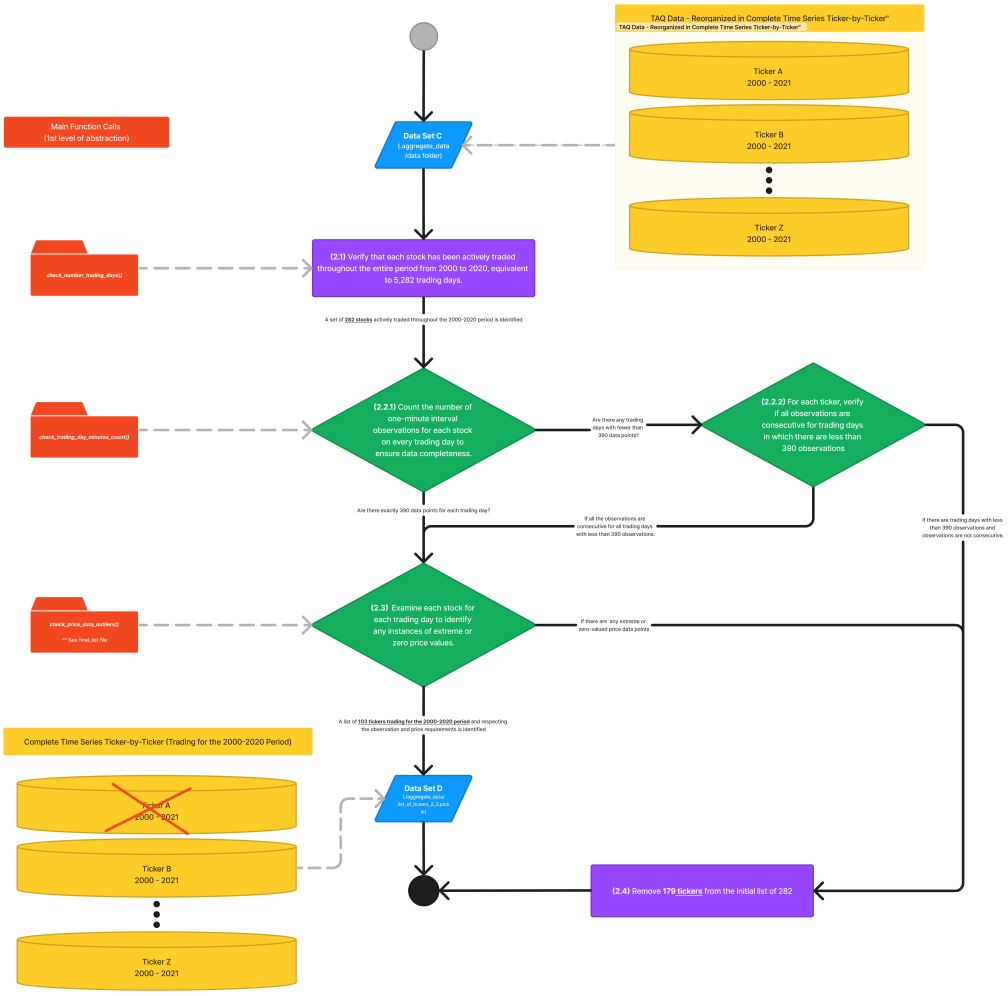

Figure 2: Step 2 - Data Validation and Cleaning

4.2.3 Step 3 - Data Completion and Model Calibration

We obtain Data Set E by retrieving the adjusted close price for the remaining stock tickers and completing Data Set D. We then calibrate four models, namely the HAR (Eq. 12), HAR-J (Eq. 13), HAR-RSV (Eq. 15), and HAR-LE (Eq. 16). We follow four substeps detailed below and in Figure 3 to perform the entire process.

Substep 1 - Adjusted Close Price and Five-Minute Subsample Estimators

We have all the necessary data points to calibrate the HAR, HAR-J, and HAR-RSV models in Data Set D. However, to calibrate the HAR-LE model, we need to compute the daily stock return components $|r_{t-1}|$ and $|r_{t-1}|I\{r_{t-1} < 0\}$ for each ticker.

We cannot use the last daily recorded trade price data from the TAQ database to construct a time series of stock returns since there is a chance of misalignment with the NYSE's official daily closing prices. Moreover, inaccuracies due to stock splits between 2000 and 2020 could lead to spurious negative returns post-split. To avoid this, we use the Python package named `yfinance` to obtain the adjusted close price time series from Yahoo Finance. We use the function `step3.compute_step_3_1()` to execute this process. We could not obtain price data for seven tickers ('ALXN', 'CERN', 'COG','CR', 'FISV', 'MDP', 'PKI') and hence exclude these tickers from Data Set D. The Data Set D is thus updated to Data Set E, which contains 96 tickers.

We then compute the subsample estimators at the five-minute interval for realized variance, bipower variation, and the realized semivariance, and we use the `step3.compute_step_3_2()` function. Additionally, we compute the daily returns for the HAR-LE model using the `step3.compute_step_3_4()` function.

Substep 2 - Trade Price Data Microstructure Noise Anomalies and Z-score Heuristic

We observe some irregularities in the information on trade prices for specific stock symbols on particular trading days. For some tickers, there are extended periods of intraday zero returns on certain trading days, indicating that the trade price data are static. Such trading days can be challenging as they result in a daily realized variance of zero, which suggests that there was no price action for the given stock. This can cause problems for model calibration because the models rely on the logarithm of realized variance estimates. When the realized variance is zero, the logarithm approaches infinity, leading to convergence difficulties in OLS fitting and model calibration failure.

To determine the extent of price stagnation, we use a heuristic formula that utilizes daily trade price Z-scores to identify potentially static trade prices on a trading day basis. We identify the maximum Z-scores (MAXZ) and the minimum Z-scores (MINZ) for each trading day and set thresholds of 0.25 and -0.25. There could be price stagnation if either MAXZ or MINZ falls within the threshold interval. This is based on the assumption that maximum and minimum Z-scores near zero might indicate no or few trade price actions in a given ticker TAQ trade price time series. We note that this situation was particularly prevalent in the year 2000. We provide further details of this analysis in Subsection 4.4. Our analysis is facilitated by the `step3.compute_step_3_3_1()` and `step3.compute_step_3_3_2()` functions.

Substep 3 - Divide Estimator Time Series in Two Trading Subperiods - 2000-2010 and 2011-2020

The subsample estimator time series and the leverage effect return estimators for the HAR-LE model are subsequently divided into two distinct subperiods — 2000 to 2010 and 2011 to 2020—using the `step3.compute_step_3_5()` function. This division is done to evaluate the performance of the models under varying market conditions and to check if they remain consistent across different trading subperiods.

The two subperiods have distinct market features and volatile shocks in stock returns. The first subperiod, from 2000 to 2010, is characterized by the Dot-com bubble and the financial crisis. The second subperiod, from 2011 to 2020, is marked by post-financial crisis market stability but is impacted by significant market turmoil at the start of 2020 due to the Covid-19 pandemic.

Substep 4 - Model Calibration

After dividing the time series into two trading subperiods, we calibrate the HAR, HAR-J, HA-RSV, and HAR-LE models with the `step3.compute_step_3_6()` function. The calibration process involves two steps: (1) a full-series model fit using OLS regression to predict the realized variance one day ahead (in-sample) and (2) a 1,000-trading day rolling window OLS estimation to predict the realized variance one day ahead (out-of-sample). The calibration process fails for seven tickers ('ANSS', 'BLK', 'CLF', 'HOLX', 'PVH', 'SWN', 'TSCO') identified as problematic during the heuristic Z-score evaluation due to the presence of trading days with a realized variance of zero. Thus, we get a final data set containing time series for 89 tickers, which are listed in the Appendix A.

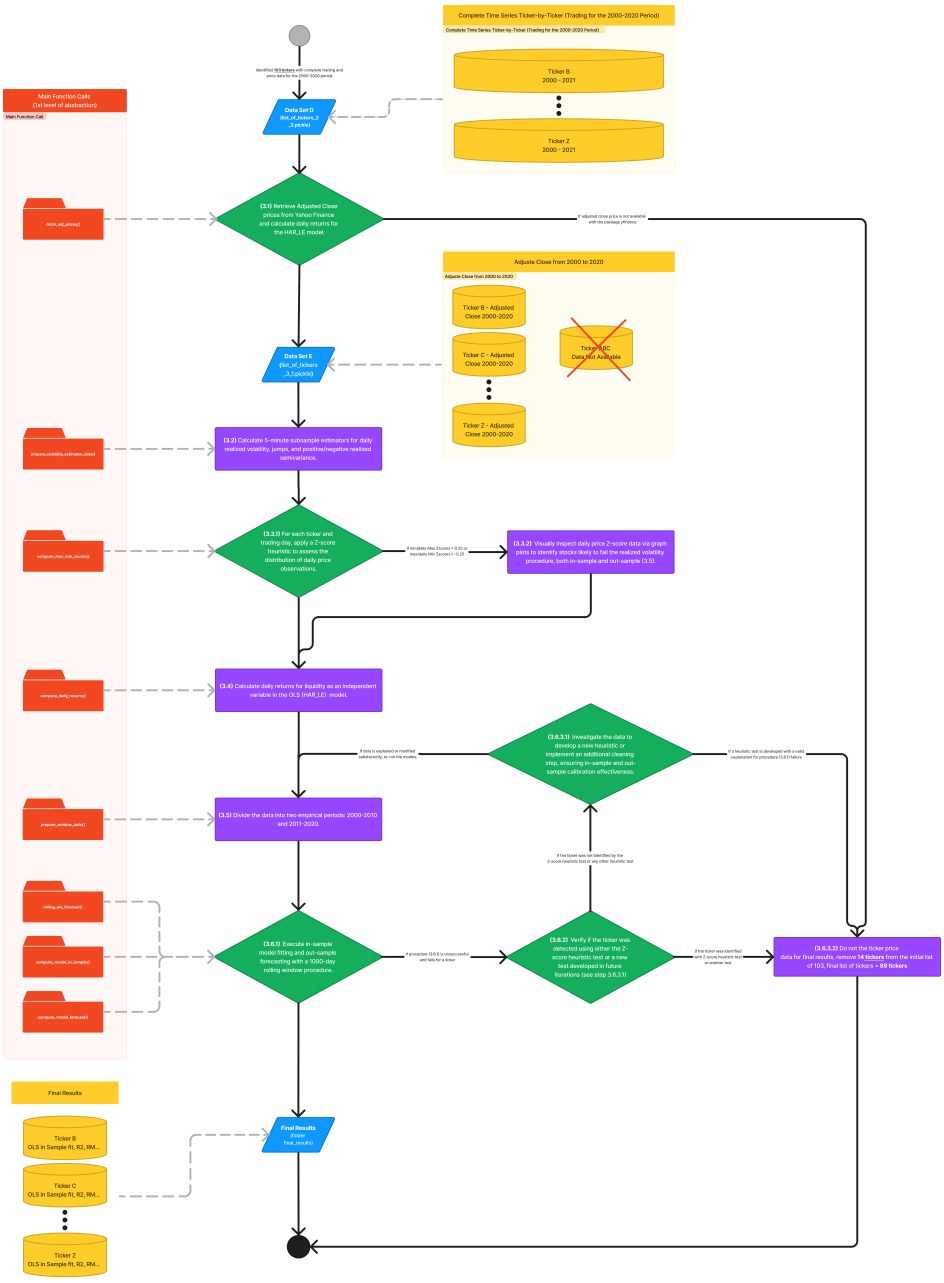

Figure 3: Step 3 - Data Completion and Model Calibration

4.3 Empirical Data Pipeline Validation

We use real-world data to validate the functionality of the Python code developed for computing realized variance subsample estimation and calibrating HAR models. We need to acquire two time series of realized volatility for this validation. First, we downloaded a realized volatility time series related to the S&P 500 Index from Professor Dacheng Xiu's website at the Booth School of Business, University of Chicago, on March 3, 2024. The URL of the data is https://dachxiu.chicagobooth.edu/. Second, we purchased the intraday "prices" of the S&P 500 Index at 1-minute intervals from FirstRateData on February 20, 2024 (more information can be found at https://firstratedata.com/). We sought clarifications from the FirstRateData support team regarding the construction of intraday prices, who explained that these prices are derived from the intraday price data of all constituent stocks within the S&P 500 Index. The FirstRateData S&P 500 Index time series starts on January 2, 2008, and ends on February 20, 2024.

We follow the same procedure outlined in Subsections 4.2.1, 4.2.2, and 4.2.3 to calculate the realized variance of the FirstRateData S&P 500 time series. We then multiply the square root of the realized variance time series by $\sqrt{252}$ to get an annualized realized volatility. We work with the annualized realized volatility to compare these results to the realized volatility time series of Dacheng Xiu.

From Figure 4, we can observe that the red time series, representing the realized volatility calculated from the realized variance 5-minute subsample estimators of the means FirstRateData S&P 500 time series, aligns with the blue line, which represents the realized variance reported by Dacheng Xiu. The green time series represents the difference between the red and blue time series, except for specific windows where volatility spikes, such as during the 2008 financial crisis and the start of 2020 during the COVID-19 pandemic, the difference remains low. These discrepancies can likely be explained by the fact that the underlying time series used by Dacheng Xiu is different than differs from the FirstRateData time series we used.

We also note from Figure 5 that the 30-day rolling correlation between the annualized realized volatility of FirstRateData's S&P 500 Index series and Dacheng Xiu's results mostly remains above zero except for a short period of volatility spike where it decreases towards zero or becomes negative. The p-values of the 30-day rolling correlation mostly remain under the significance level of 5%, as indicated by a dashed line in the figure. We can also analyze these results from the perspective of the distribution of the values of the rolling correlation and its corresponding p-values with Figure 6. The average 30-day rolling correlation is 0.4947 with

a standard deviation of 0.2184. The average p-value is 0.0767 with a standard deviation of 0.1654. These results suggest a significant and strong correlation between the two time series, indicating good alignment of the two series.

Overall, the alignment of the time series and the low difference between them indicate that the pipeline used to calculate the realized variance subsample estimators and other volatility for the HAR model calibration is reliable.

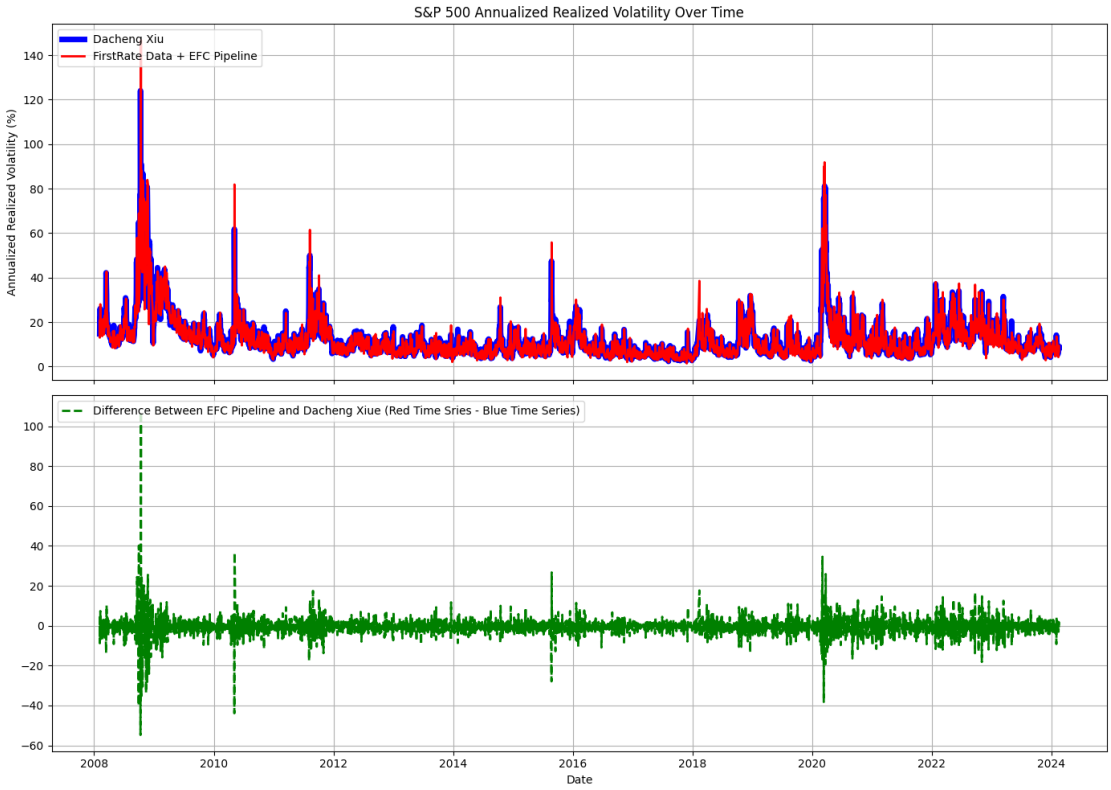

Figure 4: S&P 500 Index Realized Volatility Over Time

Note: This figure compares the annualized realized volatility of FirstRateData's S&P 500 Index series with Dacheng Xiu's results. The top graph displays both time series together, while the bottom graph shows the difference between the two.

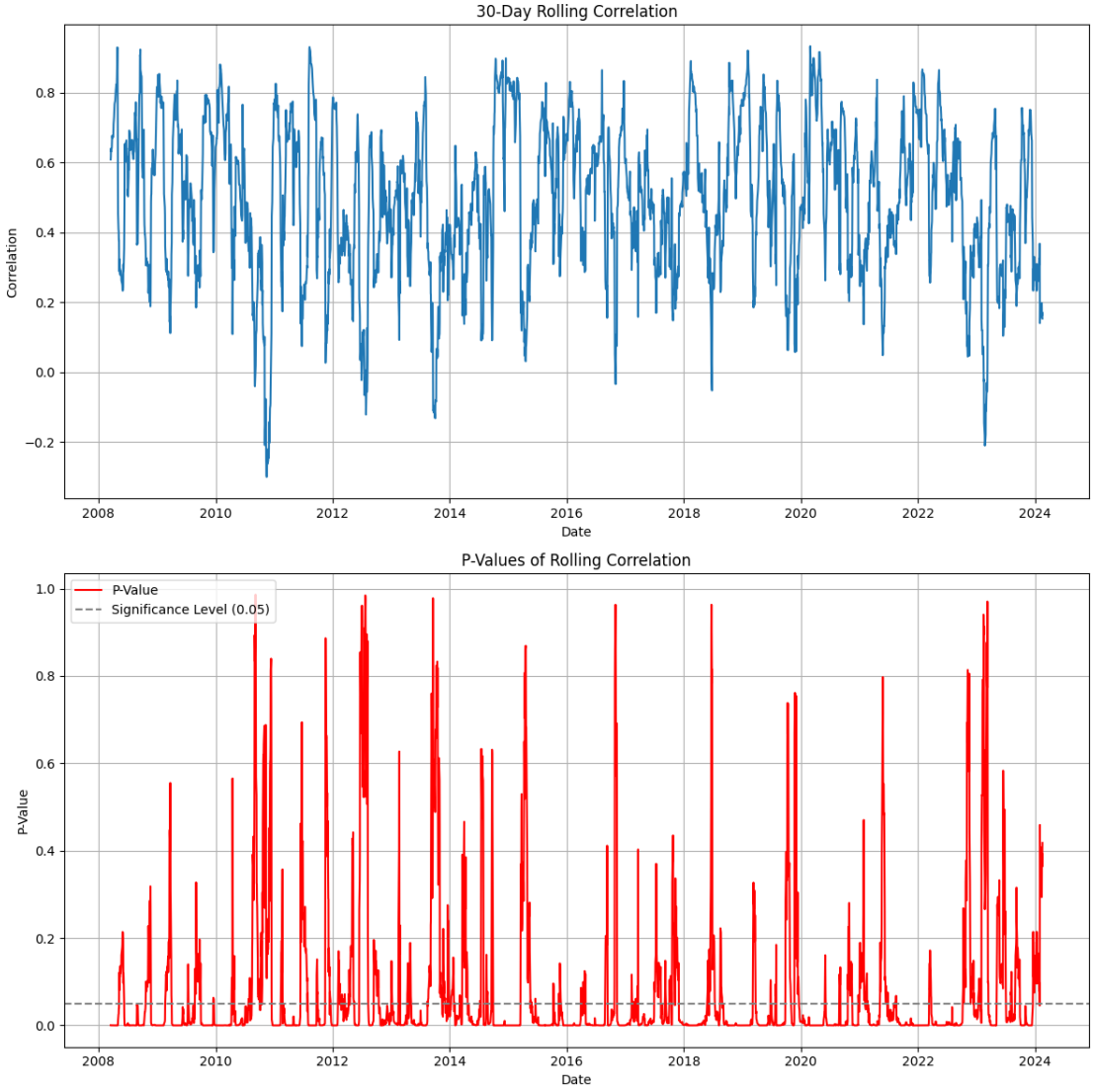

Figure 5: 30-Day Rolling Correlation and Correlation Estimator P-value Over Time

Note: The top graph shows the 30-day rolling correlation between the annualized realized volatility of FirstRateData's S&P 500 Index series and Dacheng Xiu's results of Figure 4. The bottom graph displays the p-value of the 30-day rolling correlation.

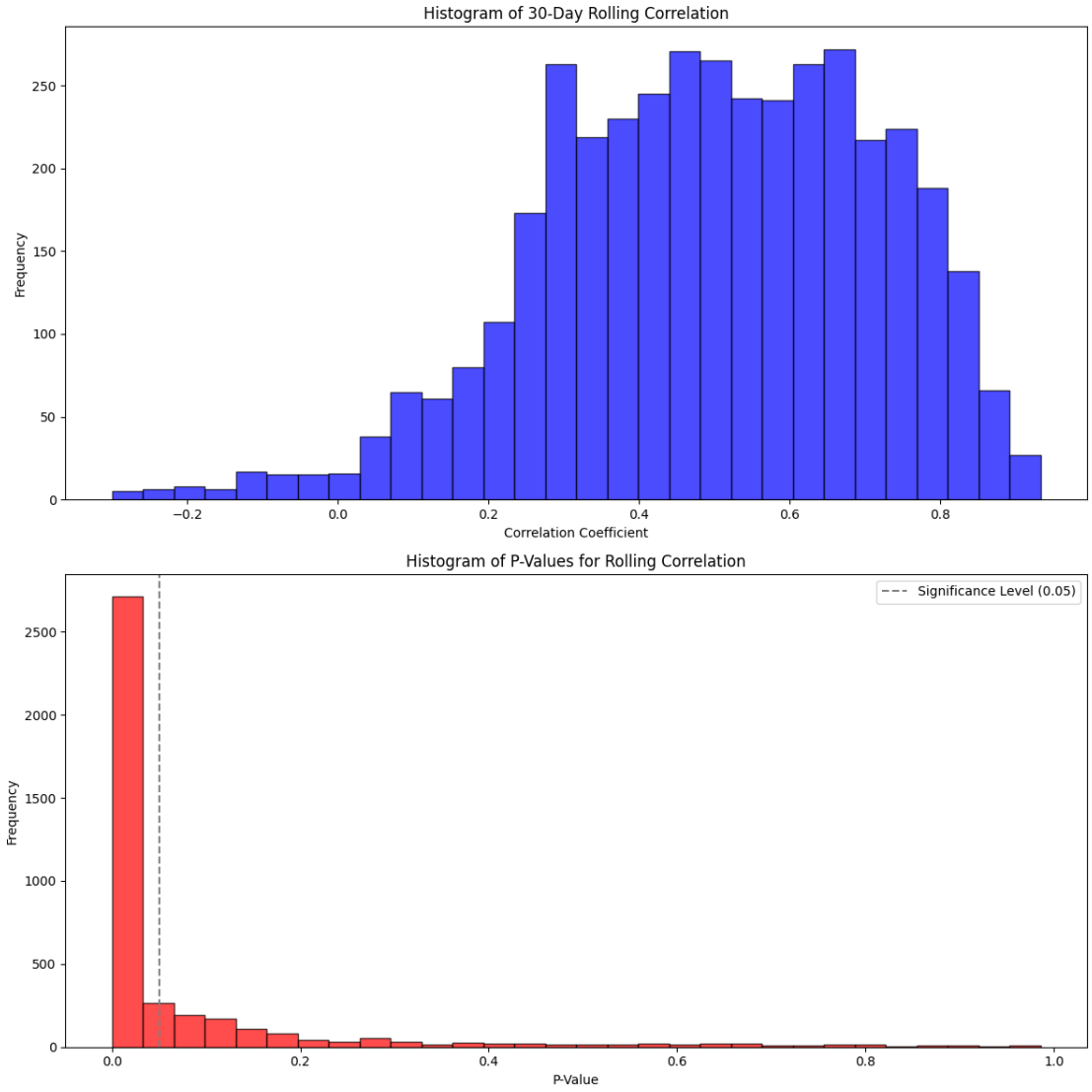

Figure 6: Correlation and P-Value Distribution

4.4 Anomalous Features of High-Frequency Data

In Section 3, we discuss the presence of market microstructure noise in trade price data from the TAQ database, which may exhibit anomalous behavior. In this subsection, we illustrate some anomalies in the raw TAQ price data with practical examples, including price deviation and price stagnation.

4.4.1 Price Deviation

Let consider the stock ticker AAPL as part of the final data set. We identify specific instances where trade prices exhibited exceptionally high Z-scores as potential anomalies. For example, Table 3 lists a trade price of \$22.18 with a Z-score of 13.88, which starkly contrasts with other trade prices on that day. Similarly, Table 4 records an open trade price of \$191.08 with a Z-score of 8.99, notably higher than subsequent prices. Figure 7 also shows examples of trading days where the daily MAXZ and MINZ exceeded the bounds of -5 and 5, respectively, for the stock ticker AAPL. This is an indication that there may be some trade prices that deviate significantly from others on the same trading days.

Table 3: TAQ Trade Price Data for AAPL on 2001-06-14

DATE	SYMBOL	itime	rtime	isize	iprice	Z Score
20010614	AAPL	15:02:00	15:01:52	100	20.13	0.14
20010614	AAPL	15:03:00	15:02:58	300	20.12	0.08
20010614	AAPL	15:04:00	15:03:57	100	20.18	0.48
20010614	AAPL	15:05:00	15:04:52	300	20.159	0.34
20010614	AAPL	15:06:00	15:05:57	600	20.04	-0.46
20010614	AAPL	15:07:00	15:06:44	5000	22.18	13.88
20010614	AAPL	15:08:00	15:07:58	200	20.15	0.28
20010614	AAPL	15:09:00	15:08:54	200	20.18	0.48

Day average price $\bar{p} = 20.1083$ Day price standard deviation σ_p 0.1493

4.4.2 Price Stagnation

Data Set D comprises 96 stock tickers after subsampling before we calibrate the HAR model. Figure 9 illustrates the distributions of the smallest values of MAXZ and MINZ for all trading days across these tickers. We identify fourteen stock tickers ('ANSS', 'BLK', 'CLF', 'HOLX',

Table 4: TAQ Trade Price Data for AAPL on 2018-07-18

DATE	SYM_ROOT	SYM_SUFFIX	itime_m	rtime_m	isize	iprice	Z-Score
20180718	AAPL		9:30:00	9:30:00	11	191.8	8.99
20180718	AAPL		9:31:00	9:31:00	35	190.92	3.71
20180718	AAPL		9:32:00	9:32:00	40	190.7	2.39
20180718	AAPL		9:33:00	9:33:00	300	190.97	4.01
20180718	AAPL		9:34:00	9:34:00	40	190.8262	3.15
20180718	AAPL		9:35:00	9:35:00	5	191	4.19
20180718	AAPL		9:36:00	9:36:00	100	190.69	2.33
20180718	AAPL		9:37:00	9:37:00	100	190.47	1.01

Day average price $\bar{p} = 190.3001$ Day price standard deviation $\sigma_p = 0.1668$

'HRL', 'IFF', 'IP', 'PCG', 'PPL', 'PVH', 'SNV', 'SWK', 'SWN', 'TSCO') where at least one trading day for each exhibited MAXZ or MINZ values within the Z-score range of -0.25 to 0.25. This range may indicate the presence of stagnant prices on at least one trading day within these time series. It is worth noting that having prices within this particular range does not necessarily imply that all trade prices are constant for at least one trading day, leading to a daily realized variance of zero. Seven of the 14 stock tickers flagged with the MAXZ or MINZ encountered calibration issues as detailed in Subsection 4.2.3. The calibration failure occurs due to at least one daily realized variance subsample estimator with a value of zero and a logarithm of infinity.

The MAXZ and MINZ values over time for the AAPL stock ticker show no signs of price stagnation in Figure 7. In contrast, Figure 8 indicates the ANSS stock ticker potential price stagnation on certain trading days. Notably, at the end of the trading year 2000, a clear pattern of price stagnation is evident as both MAXZ and MINZ values gradually converge towards zero. Unlike AAPL, where no similar trend is observed, ANSS displays trading days with a realized variance subsample estimator of zero, leading to calibration failures in the HAR model.

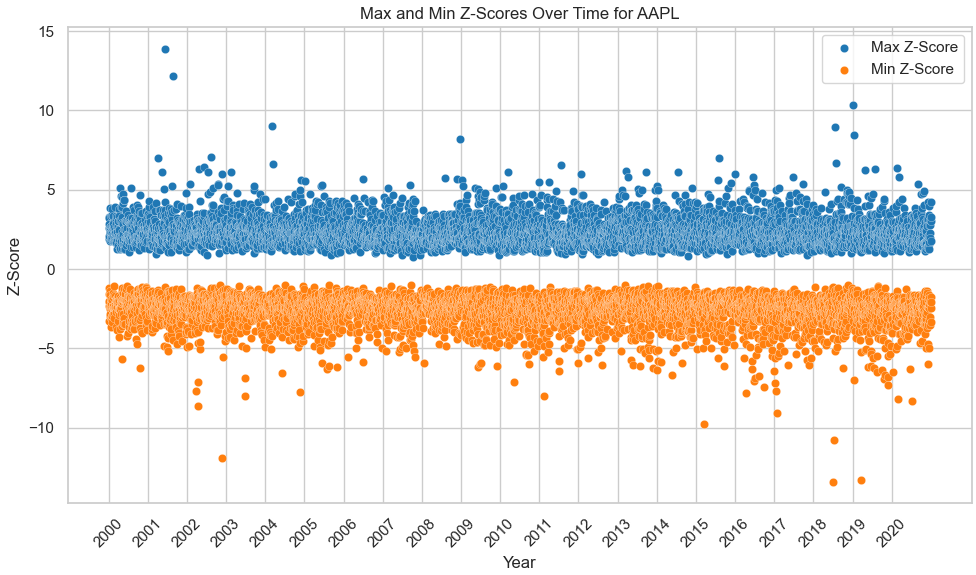

Figure 7: Daily Maximum and Minimum Z-Scores Over Time for AAPL

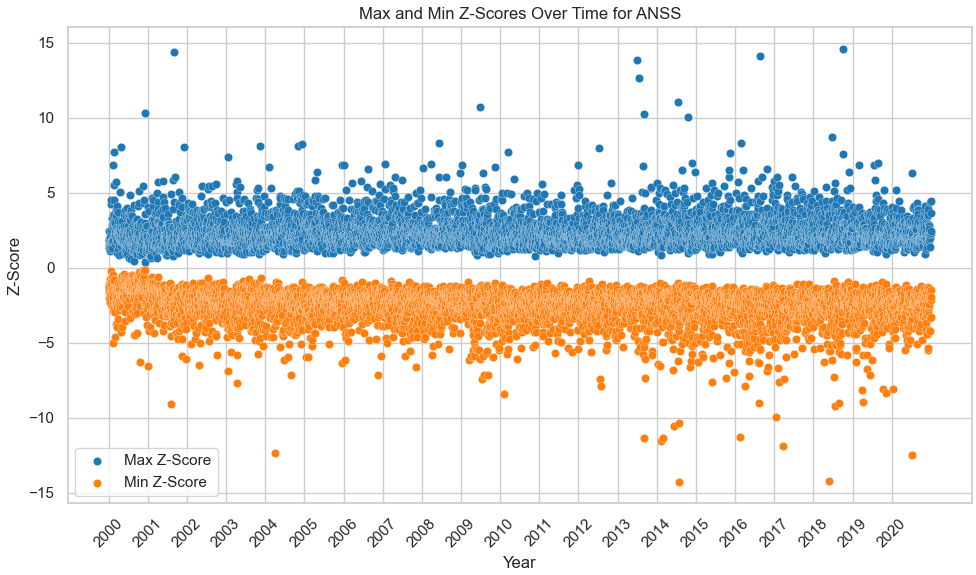

Figure 8: Daily Maximum and Minimum Z-Scores Over Time for ANSS

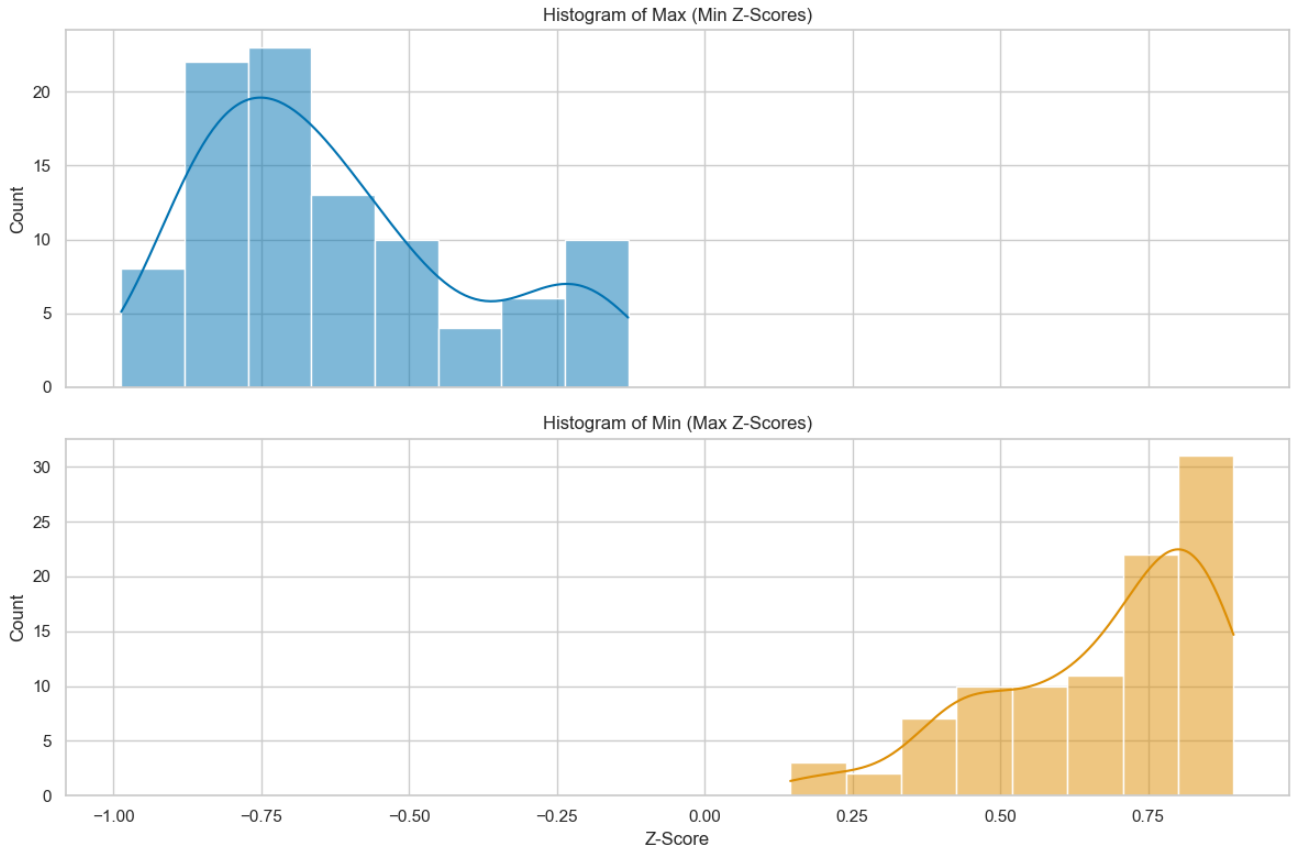

Figure 9: Analysis of Heuristic Maximum and Minimum Z-Scores

5 Empirical Analysis

5.1 Data Preliminary Analysis

Table 5 presents summary statistics of the estimates for realized variance, bipower variation and realized semivariance obtained from the intraday trade price data for the FirstRateData S&P 500 time series and from the five-minute subsample estimators calculated from the NYSE TAQ data using the trade price data from individual stocks.

Panel A of Table 5 shows that individual firms' average daily realized variance is 0.000621 (equivalent to an annualized realized variance of 15.64%) between 2000 and 2010. The average annualized realized volatility, obtained by taking the square root of the daily realized volatility and then averaging across stocks[2], is 32.06%, which aligns with the results presented by Patton and Sheppard (2015) (33.2%, based on a dataset of 105 stocks from 1997 to 2008) and Lyócsa and Todorova (2020) (31.17%, from a sample of 431 stocks spanning 2007 to 2016).

Panel B reveals that the average daily realized variance for the S&P 500 Index during the trading period from 2011 to 2020 is 0.00060 (equivalent to an annualized realized variance of 1.52%). The annualized daily realized volatility for the S&P 500 Index is 9.86%, below the figure reported by Patton and Sheppard (2015) of 17.1%. The mean daily realized variance for individual stock symbols is 0.000304 (annualized realized variance of 7.66%), and the average annualized realized volatility for individual stocks is 22.71%, which is relatively lower compared to the results documented by Patton and Sheppard (2015) and Lyócsa and Todorova (2020).

Panel B's trading period only starts on January 1, 2011, and does not overlap with the 2007-2008 financial crisis compared to the studies of cited authors. This might explain why the annualized realized variance we observe is lower for the S&P 500 and individual firms. Also, Panel A covers the trading period during the 2008 financial crisis, a time of unusually high levels of stock market volatility. Research on stock market volatility in the United States, the United Kingdom, and Japan suggests that the volatility observed during the 2008 crisis was short-lived (Schwert, 2011). This observation may help explain the subsequent decrease in average realized variance during the trading period from 2011 to 2020, which we note in Panel B, compared to 2000-2010.

[2]The daily realized volatility values are not shown in Table 5.

The "Max" column in Table 5 reveals an intriguing pattern concerning the average value of the estimators (RV, BV, RSV^+, and RSV^-) presented in both Panels A and B. This column indicates that the maximum values for all estimators are much higher than those at the 95^{th} percentile. These observations align with the discussions on price deviations in HFD previously detailed in Subsection 4.4.1. Such substantial disparities in realized variance underscore the importance of employing a subsample estimator at the five-minute interval. This approach not only smooths the final estimators but also minimizes the impact of these extreme values on the overall results. In contrast, the "Min" column shows no notable difference between the minimum average estimator values and those at the 5^{th} percentile.

Table 5: Data Summary Statistics

	S&P 500	Mean	$Q_{0.05}$	Median	$Q_{0.95}$	Min	Max
Panel A: 2000-2010							
Averages							
RV	-	0.621	0.059	0.268	2.081	0.002	461.172
BV	-	0.521	0.051	0.232	1.801	0.000	281.088
RSV^+	-	0.304	0.027	0.130	1.050	0.000	209.013
RSV^-	-	0.317	0.027	0.129	1.037	0.000	461.027
Autocorrelations (1 lag)							
RV	-	0.3941	0.0083	0.4745	0.7104	0.0018	0.7737
BV	-	0.7149	0.4613	0.7330	0.8619	0.2833	0.9056
RSV^+	-	0.4275	0.0459	0.4850	0.6357	0.0093	0.6875
RSV^-	-	0.3609	0.0017	0.4352	0.6795	-0.0002	0.7505
Panel B: 2011-2020							
Averages							
RV	0.060	0.304	0.040	0.134	0.934	0.006	193.174
BV	0.056	0.274	0.036	0.122	0.848	0.006	107.181
RSV^+	0.030	0.154	0.019	0.066	0.473	0.003	185.845
RSV^-	0.030	0.150	0.018	0.065	0.471	0.001	32.891
Autocorrelations (1 lag)							
RV	0.8141	0.6660	0.3413	0.7066	0.8392	0.1083	0.8826
BV	0.8240	0.7149	0.4613	0.7330	0.8619	0.2833	0.9056
RSV^+	0.7859	0.6013	0.1950	0.6507	0.7874	0.0556	0.8538
RSV^-	0.7696	0.6526	0.4152	0.6777	0.7989	0.1036	0.8667

Note: The information provided in the S&P 500 column pertains to the average daily values of realized variance (RV), bipower variation (BV), positive semivariance (RSV^+), and negative semivariance (RSV^-) estimators, specifically within the *Averages* subpanels and is solely computed based on the S&P 500 time series. The averages shown in the neighboring columns (Mean, $Q_{0.05}$, Median, $Q_{0.95}$, Min, and Max) are derived from the entire dataset consisting of 89 stock ticker time series. This also holds for the *Autocorrelations* subpanels, where the autocorrelations indicate the lag-1 autocorrelation and only present the raw autocorrelation value for the S&P 500 time series in the S&P 500 column, with the average autocorrelation for the adjacent columns being calculated across the complete dataset of 89 stock tickers. Furthermore, all values in the *Averages* subpanels are scaled by a factor of 1,000.

5.2 In-Sample Forecast Performance

We calibrate HAR models (HAR, HAR-J, HAR-RSV and HAR-LE) in-sample using OLS regression to predict the one-day ahead logarithmic realized variance for the S&P 500 time series (FirstRate Data Set) and each of the 89 individual stocks (TAQ Data Set).

Every regression model employs the Newey-West estimator to address autocorrelation and heteroskedasticity in the error terms. The choice of lags for the Newey-West estimator is based on the work of Newey and West (1994) where the highest lag to include in the kernel, L, is calculated with the following formula, $L = \left\lfloor 4 \left(\frac{T}{100}\right)^{\frac{2}{9}} \right\rfloor$. The Bartlett kernel is used with Newey-West's automatic bandwidth selection method.

The OLS regression coefficients for the fitted models for the S&P 500 time series during the trading period of 2011-2020 are presented in Table 6. The results for the 89 stocks in the trading periods 2000-2010 and 2011-2020 are aggregated and presented in Table 7.

5.2.1 S&P 500 Index

The following analysis examines the OLS regression results of the S&P 500 time series. As seen in Table 6, the OLS coefficients are significant at the 1% level for each regressed log lagged realized variance regression variable of the HAR model, namely $\ln RV^d_{t-1}$, $\ln RV^w_{t-1}$, and $\ln RV^m_{t-1}$. The extended models, which consider the volatility asymmetry components, HAR-J, HAR-RSV, and HAR-LE, also display significant OLS coefficients for the same three HAR model log lagged realized variance regression variables.

We observe that the importance of the log realized variance regressors decreases as the time horizon increases from daily to weekly to monthly regressions for the S&P 500 time series. This can be seen from the OLS coefficient for $\ln RV^d_{t-1}$ being larger than that for $\ln RV^w_{t-1}$ and $\ln RV^m_{t-1}$, and the coefficient for RV^w_{t-1} being higher than that for RV^m_{t-1}. These findings are consistent with previous studies by Corsi (2009), Patton and Sheppard (2015), and Andersen et al. (2007), despite differences in the analysis time frame and the S&P 500 time series used. Corsi (2009) covers the period from 1990 to 2007, Patton and Sheppard (2015) cover the period from 1997 to 2008, and Andersen et al. (2007) cover the period from 1990 to 2002.

We note that during the in-sample calibration of the HAR-J model, the OLS coefficient of $\ln(J_{t-1}+1)$ is negative, which has also been reported in Andersen et al. (2007). Our results reveal that the jump coefficient is not significant and does not contribute to improving the in-sample performance. This can be seen in the adjusted R-squared of both the HAR and HAR-J models, which show only a slight difference and no increase for the HAR-J model

Table 6: In-sample One-Day Ahead Realized Variance Model Parameters Evaluation For the S&P 500 (2011-2020)

Trading Window Period	2011-2020			
	HAR	HAR-J	HAR-RSV	HAR-LE
Intercept	-0.882^{**}	-0.833^{**}	-0.543^{**}	-1.659^{**}
$\ln RV_{t-1}$	0.516^{**}	0.519^{**}		0.433^{**}
$\ln RV^{w}_{t-1}$	0.265^{**}	0.265^{**}	0.299^{**}	0.270^{**}
$\ln RV^{m}_{t-1}$	0.141^{**}	0.141^{**}	0.139^{**}	0.153^{**}
$\ln(J_{t-1}+1)$		-773.841		
$\ln RSV^{+}_{t}$			0.156^{**}	
$\ln RSV^{-}_{t}$			0.327^{**}	
$\lvert r_{t-1}\rvert$				5.240^{*}
$\lvert r_{t-1}\rvert I\{r_{t-1}<0\}$				19.806^{**}
Adjusted R^2	69.091	69.092	69.157	70.317

Note: *, and ** denote rejections of null hypothesis at 5% and 1% significance levels, respectively

when $\ln(J_{t-1}+1)$ is added. Maki and Ota (2021) found similar results, showing that the coefficient of $\ln(J_{t-1}+1)$ is both negative and statistically insignificant for the Japanese market for the Nikkei 225 time series that spans from 2001 to 2019.

The R-squared values of the HAR-RSV and HAR-LE models suggest that they perform better in-sample to predict one-day ahead forecasts than the HAR and HAR-J models. In particular, the HAR-RSV model shows that the coefficients of $\ln RSV^{+}_{t}$ and $\ln RSV^{-}_{t}$ are significant at the 1% level. The coefficient for the log negative realized semivariance term is higher than that of the log positive realized semivariance, indicating that negative semivariance could be a better predictor of future volatility than positive realized semivariance for the S&P 500 Index.

Likewise, the HAR-LE model's leverage effect has shown the best in-sample performance compared to other fitted models. Both $|r_{t-1}|$ and $|r_{t-1}|I\{r_{t-1}<0\}$ have a significant OLS coefficient, although only the coefficient of $|r_{t-1}|I\{r_{t-1}<0\}$ is significant at the 1% level. Furthermore, the coefficient for negative past returns is greater than for positive past returns, suggesting a higher forecasting power for future volatility. This finding is consistent with what is discussed in detail bu Corsi et al. (2008).

5.2.2 Individual Firms

We examine the in-sample OLS regression results for the different calibrated models for the 89 final stock tickers during the two trading periods from 2000 to 2010 and 2011 to 2020. According to Table 7, we find that, across the four models and both trading periods, the average coefficient for $\ln RV^d_{t-1}$ is higher than that of $\ln RV^w_{t-1}$ and $\ln RV^m_{t-1}$ and the average coefficient for $\ln RV^w_{t-1}$ tends to be lower than that of $\ln RV^m_{t-1}$. This is different from the findings of Lyócsa and Todorova (2020) and Andersen et al. (2007) for the HAR model, where the mean regression coefficient in-sample for RV^w_{t-1} is higher than that of $\ln RV^m_{t-1}$ in one-day ahead forecasts. According to Lyócsa and Todorova (2020), the average in-sample regression coefficient for $\ln RV^m_{t-1}$ only becomes more significant than that of $\ln RV^w_{t-1}$ for longer-term forecasts, specifically five-days ahead, and even exceeds that of $\ln RV^d_{t-1}$ for 22-days ahead regression. Similar results are also reported by Patton and Sheppard (2015) for the HAR model, where the average coefficient for $\ln RV^m_{t-1}$ is only higher than that of RV^w_{t-1} for the 22-days ahead regression for in-sample data. We notice that in almost half of the analyzed stock symbols, the coefficient associated with $\ln RV^w_{t-1}$ was higher than that of $\ln RV^m_{t-1}$. Specifically, during the trading period 2000-2010, 42 coefficients for $\ln RV^w_{t-1}$ out of 89 for HAR, 44 for HAR-J, 44 for HAR-RSV, and 41 for HAR-LE exhibit this characteristic, while 42 for HAR, 45 for HAR-J, 43 for HAR-RSV, and 38 for HAR-LE show it during the trading period of 2011-2020. Notably, not all stock tickers showed a regression coefficient that is higher for $\ln RV^m_{t-1}$ than for $\ln RV^w_{t-1}$. Instead, the overall coefficient average indicates this tendency. Lastly, Figures 10 and 11 show that the three variables of lagged log realized variance ($\ln RV^d_{t-1}$, $\ln RV^w_{t-1}$, and $\ln RV^m_{t-1}$) are significant at the 1% level across the four models tested during both trading periods from 2000 to 2010 and 2011 to 2020.

We observe that the coefficients of $\ln(J_{t-1}+1)$ in the HAR-J model fitted in-sample for the individual firms are similar to the ones obtained from the HAR-J model fitted to the S&P 500 time series. However, unlike the S&P 500 time series, the average adjusted R-squared increases in both trading window periods when compared to the R-squared of the standard HAR model. This suggests that the jump component might play a role in improving the performance in-sample for individual firms and reducing the impact of lagged realized volatility, as pointed out by Andersen et al. (2007). Analyzing the period from 2000 to 2010, we see that out of the 89 ticker symbols, the HAR-J model has OLS regression coefficients for $\ln(J_{t-1}+1)$ that are significant at the 5% level for 74 stocks. Similarly, during the trading period from 2011 to 2020, this significance is observed for 41 stock tickers. The jump component shows significance for almost half of the stock tickers examined.

Table 7: In-sample Realized Variance Model Parameters Evaluation For Firms

Trading Window Period	2000-2010				2011-2020			
	HAR	HAR-J	HAR-RSV	HAR-LE	HAR	HAR-J	HAR-RSV	HAR-LE
Intercept	-0.671	-0.382	-0.398	-0.948	-1.017	-0.860	-0.704	-1.422
	(0.279)	(0.259)	(0.256)	(0.315)	(0.324)	(0.305)	(0.324)	(0.383)
$\ln RV_{t-1}$	0.355	0.382		0.325	0.421	0.444		0.370
	(0.045)	(0.050)		(0.044)	(0.03)	(0.035)		(0.032)
$\ln RV^{w}_{t-1}$	0.281	0.282	0.277	0.282	0.228	0.227	0.224	0.231
	(0.029)	(0.051)	(0.054)	(0.051)	(0.04)	(0.040)	(0.041)	(0.039)
$\ln RV^{m}_{t-1}$	0.290	0.283	0.287	0.292	0.242	0.237	0.238	0.251
	(0.064)	(0.066)	(0.064)	(0.065)	(0.038)	(0.034)	(0.038)	(0.038)
$\ln(J_{t-1}+1)$		-215.051				-494.179		
		(197.800)				(447.519)		
$\ln RSV^{+}_{t}$			0.136				0.177	
			(0.051)				(0.023)	
$\ln RSV^{-}_{t}$			0.228				0.252	
			(0.032)				(0.027)	
$\lvert r_{t-1}\rvert$				1.207				2.381
				(1.211)				(1.75)
$\lvert r_{t-1}\rvert I\{r_{t-1}<0\}$				4.572				6.430
				(1.811)				(2.659)
Average Adjusted R^2	69.3	69.7	69.5	69.7	60.4	60.6	60.5	61.0
	(7.79)	(7.54)	(7.74)	(7.73)	(8.01)	(7.92)	(7.97)	(7.94)

Note: The values in the table correspond to average coefficients across the final data set of 89 stock tickers. The values in parentheses are standard deviations of the coefficients calculated across all 89 stock tickers.

We notice that negative semivariance parameter, $\ln RSV_{t-1}^{-}$, has a more significant effect on $\ln RV_t$ than positive semivariance, $\ln RSV_{t-1}^{+}$ since, during both trading periods, the regression coefficients for the negative semivariance realized are higher than those for $\ln RSV_{t-1}^{+}$ (0.228 vs. 0.136 for 2000-2010 and 0.252 vs. 0.177 for 2011-2020). These findings are consistent with the conclusions of Maki and Ota (2021) and Patton and Sheppard (2015), which emphasize the importance of negative semivariance in forecasting future realized variance and volatility. The p-values of the regression coefficient for $\ln RSV_{t-1}^{+}$ and $\ln RSV_{t-1}^{-}$ in the trading periods 2000-2010 and 2011-2020 are primarily significant at the 1% level for most stock tickers, as shown in Figures 12 and 13.

Table 7 also shows that the average R-squared result obtained from the HAR-LE model fitting is higher than that of other models, as observed in Subsection 5.2.1. However, it is worth noting that when analyzing individual firms, we get a different perspective on the predictive power of positive past returns. The coefficient of positive past returns, $|r_{t-1}|$, is significant for the S&P 500, but it is only significant for 27 out of 89 stock tickers during the trading period of 2000-2010 and 43 during 2011-2020, as shown in Figures 12 and 13. The distribution of the adjusted R-squared of all 89 tickers across models and time periods are available in Figures 14 and 15.

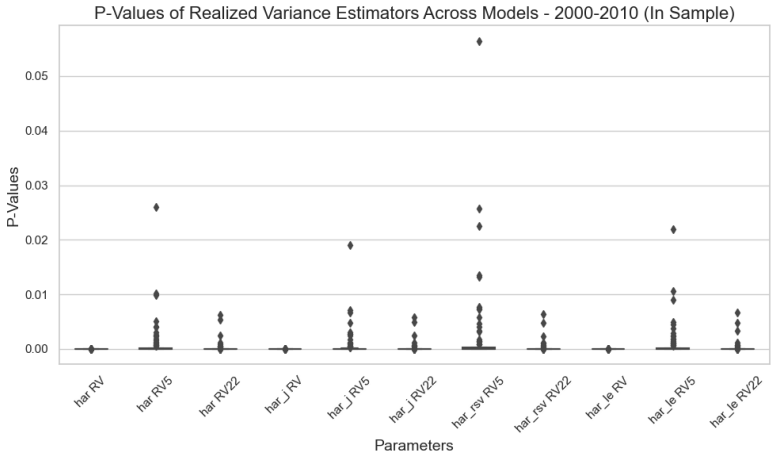

Figure 10: P-Values of Realized Variance Estimators Across Models - 2000-2010 (In-Sample)

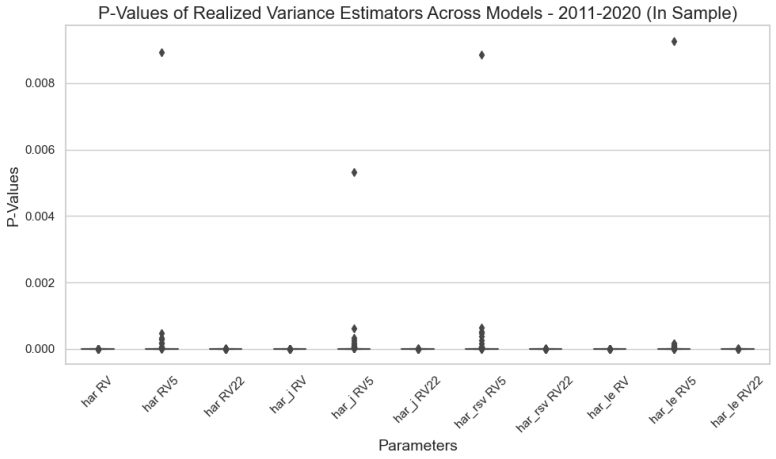

Figure 11: P-Values of Realized Variance Estimators Across Models - 2011-2020 (In-Sample)

Note: Figures 10 and 11 show, for the 89 stock tickers, the distribution of the p-values of the parameters $\ln RV^d_{t-1}$ (identified with the symbol "RV" on the figures), $\ln RV^w_{t-1}$ (RV5), and $\ln RV^m_{t-1}$ (RV22), which are common among the four tested models, HAR, HAR-j, HAR-RSV, and HAR-LE. For example, the column "har RV5" corresponds to the distribution of p-values of the parameter $\ln RV^w_{t-1}$ for the HAR model.

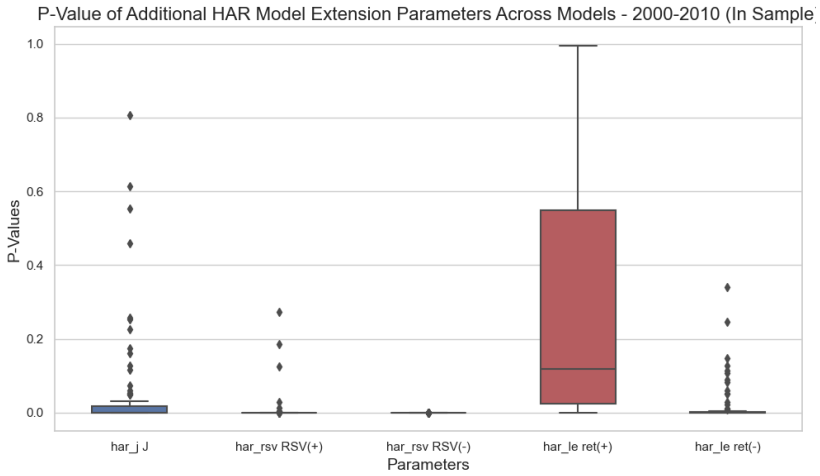

Figure 12: P-Value of Additional HAR Model Extension Parameters Across Models - 2000-2010 (In-Sample)

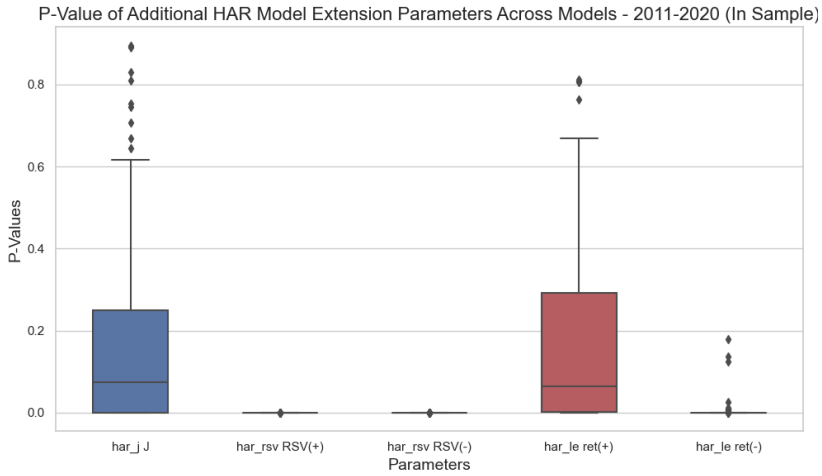

Figure 13: P-Value of Additional HAR Model Extension Parameters Across Models - 2000-2010 (In-Sample)

Note: Figures 12 and 13 show, for the 89 stock tickers, the distribution of the p-values of the parameters $\ln(J_{t-1}+1)$ (identified with the symbol "J" on the figures), $\ln RSV^{+}_{t-1}$ (RSV(+)), $\ln RSV^{-}_{t-1}$ (RSV(-)), $|r_{t-1}|$ (ret(+)), and $|r_{t-1}|I\{r_{t-1}<0\}$ (ret(-)) which extends the HAR models in the models HAR-J, HAR-RSV, and HAR-LE. For example, the column "har_rsv RSV(+)" corresponds to the distribution of p-values of the parameter $\ln RSV^{+}_{t-1}$ for the HAR-RSV model.

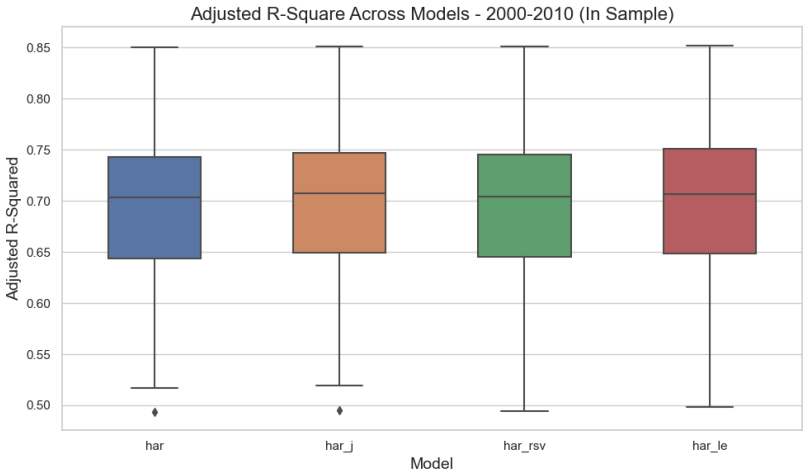

Figure 14: Distribution of Adjusted R-Squared Across Models - 2000-2010 (In-Sample)

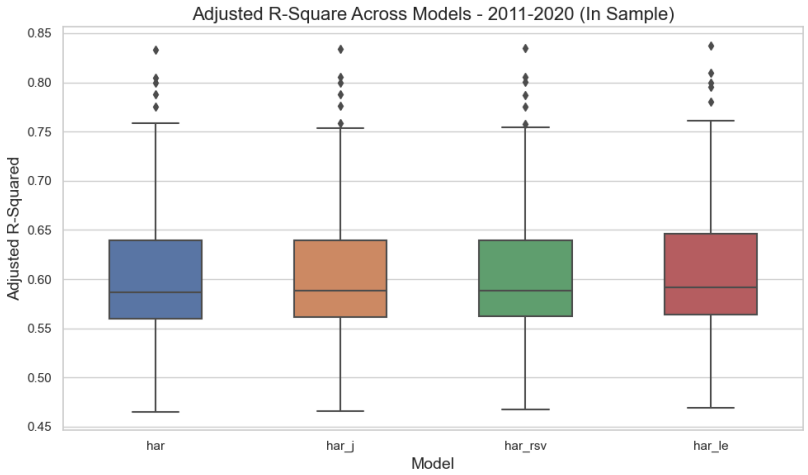

Figure 15: Distribution of Adjusted R-Squared Across Models - 2011-2020 (In-Sample)

5.3 Out-of-Sample Forecasting Performance

5.3.1 Forecasting Procedure

We use a rolling window approach to evaluate the forecast performance of the different models on the S&P 500 time series (FirstRate Data Set) and each of the 89 individual stocks (TAQ Data Set). Specifically, we calibrate the HAR, HAR-J, HAR-RSV, and HAR-LE models using training sets that comprised 1,000 trading days. Our objective was to forecast one-day ahead of the out-of-sample log variance realized and assess the models' forecast accuracy. Figure 16 depicts this rolling window method.

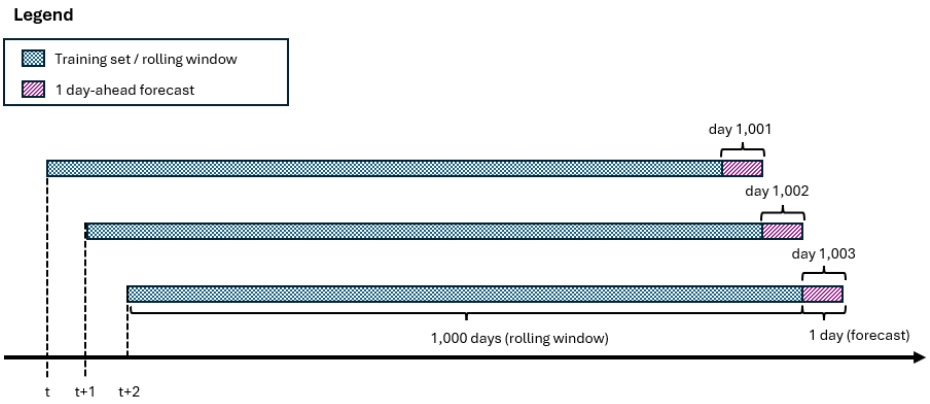

Figure 16: Illustration of the Rolling Window Approach

5.3.2 Forecast Performance Evaluation

We evaluate the accuracy of a forecasting model using two different loss functions: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). We first examine how each model performs based on the raw measurement of the loss function. After that, we conduct a Diebold-Mariano test (Diebold and Mariano, 2002) with the adjustment of Harvey et al. (1997) to determine the significance of the accuracy differences between each model extension (HAR-J, HAR-RSV, and HAR-LE) and the HAR model. We then compare each extension against the other.

The RMSE and MAE loss functions are defined as follows:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\ln RV_t} - \ln RV_t\right)^2}, \text{ and} \tag{19}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\widehat{\ln RV_t} - \ln RV_t\right|, \tag{20}$$

where $\widehat{\ln RV_t}$ represents the forecasted log realized variance, $\ln RV_t$ is the actual log realized variance and T the number of predicted days. We choose RMSE and MAE due to their differing sensitivities to outliers, with RMSE placing a higher penalty on large forecast errors.

5.3.3 S&P 500 Index

Table 8 presents the out-of-sample performance for the S&P 500 time series, detailing the RMSE and MAE values. The RMSE values are consistently higher than the MAE values across all models. The HAR-J model has a slightly lower RMSE than the HAR and HAR-RSV models. However, the differences in MAE among these models are minimal. Table 9 displays the Diebold-Mariano statistics and supports these results. Regarding RMSE, the HAR-J's greater accuracy compared to HAR is not statistically significant, and the lower accuracy of the HAR-RSV compared to HAR is only significant at the 10% level. Regarding MAE, neither the HAR-J nor HAR-RSV models have forecast accuracy significantly different than that of the HAR model.

Table 8: RMSE and MAE of Out-of-Sample Log Realized Variance Across the 2010-2020 Trading Period

	HAR	HAR-J	HAR-RSV	HAR-LE
RMSE	0.6186	0.6179	0.6194	0.6023
MAE	0.4915	0.4914	0.4915	0.4786

Table 8 also shows that the HAR-LE model stands out among competing models by exhibiting the lowest RMSE and MAE values, suggesting superior predictive accuracy compared to the benchmark HAR model and other competing models. Table 9 demonstrates that the accuracy of the HAR-LE model is significantly greater at the 5% level compared to the HAR

model, both in terms of RMSE and MAE. These results are consistent with findings from previous studies such as those by Maki and Ota (2021) and Corsi and Renò (2009), which demonstrate that models incorporating the leverage effect perform better out-of-sample for equity indices such as the Nikkei 225 and S&P 500, respectively. The forecasted realized variance of the HAR-LE model compared to the realized variance of the S&P 500 is depicted in Figure 17.

Table 9: Diebold-Mariano Statistics for RMSE and MAE Loss Functions and Model Accuracy Significance Comparison

Model Comparison	**HAR-J**	**HAR-RSV**	**HAR-LE**
RMSE			
Benchmark			
HAR	1.469	-1.666^{*}	1.953^{**}
HAR-J	–	-2.158^{**}	1.801^{*}
HAR-RSV	–	–	2.644^{***}
MAE			
Benchmark			
HAR	1.208	-0.645	2.155^{**}
HAR-J	–	-1.071	2.006^{**}
HAR-RSV	–	–	2.452^{**}

Note: This table shows Diebold-Mariano statistics comparing forecast errors across different models using RMSE and MAE loss functions. A benchmark model's forecast error is compared to another model's error (column entries) to compute these statistics. A positive significant Diebold-Mariano statistic indicates that the benchmark model has significantly higher error (and thus lower accuracy) than the tested model; a negative significant statistic indicates the opposite. For instance, when HAR is benchmarked against HAR-LE with an RMSE measure, the statistic is 1.953, which is positive and significant at the 5% level, indicating lower HAR accuracy than HAR-LE. Significance levels are indicated by * (10%), ** (5%), and *** (1%).

5.3.4 Individual Firms

We assess the forecast performance of various volatility models for 89 stock tickers over two distinct training periods: 2000-2010 and 2011-2020. The evaluation of RMSE and MAE metrics is visually detailed in box plots shown in Figures 18, 19, 22, and 23. These plots generally indicate that the loss function results for models HAR-J, HAR-RSV, and HAR-LE are more favorable than those from the standard HAR model, displaying lower values for

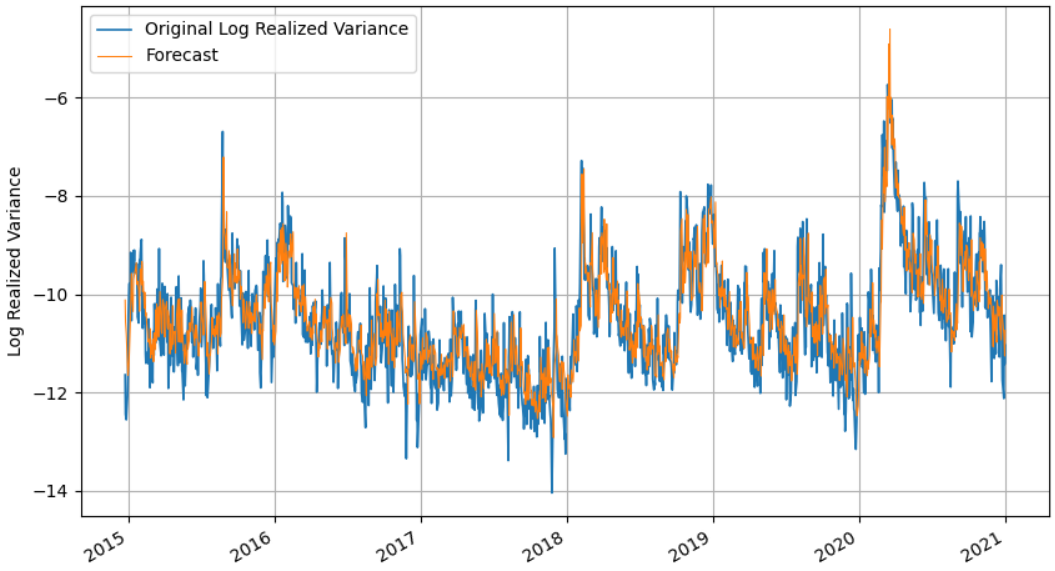

Figure 17: S&P 500 Index One-Day Ahead Forecast Logarithmic Realized Variance for Model HAR-LE vs Empirical Values

both RMSE and MAE across both periods.

Further insights are provided through count matrix heatmaps, as depicted in Figures 20, 21, 24, and 25. These heatmaps effectively illustrate the comparative performance of each model (identified on the y-axis of the heatmaps) by showing the number of stocks, out of the total 89, for which each model recorded the lowest to the highest RMSE and MAE values (displayed on the x-axis of the heatmaps from 0, the lowest loss function value, to 3, the highest loss function value). For instance, in Figure 24, the HAR model is shown to have the highest RMSE values for 78 out of the 89 stocks during 2000-2010, while the HAR-LE model achieved the lowest RMSE values for 60 stocks in the same period.

Consistent patterns emerge from the heatmaps across all metrics and both training periods. The standard HAR model, lacking a volatility asymmetry component, consistently shows the highest loss function value count (3 on the x-axis). In contrast, the HAR-LE model stands out with the best performance, having the lowest loss function results (0 on the x-axis) for most stocks. The HAR-J model, which sits between the standard HAR and HAR-LE models in terms of performance, generally outperforms the HAR-RSV model. Specifically, the HAR-

J model exhibits a greater frequency of the lowest (0 on the x-axis) and second-lowest (1 on the x-axis) loss function values for RMSE and MAE. In contrast, the HAR-RSV model more frequently records the second-highest loss function values (2 on the x-axis).

While the box plots and count matrix heatmaps suggest a certain hierarchy in model performance, the Diebold-Mariano test results add nuance to these findings. In Figures 26 and 27, we observe that during the 2000-2010 trading period, the HAR-RSV model shows significantly better accuracy than the HAR model for a substantial proportion of the stocks under both RMSE and MAE metrics. However, the HAR-J and HAR-LE models do not generally demonstrate a statistically significant improvement in accuracy over the HAR model. Moreover, when comparing the HAR-RSV model against the HAR-J and HAR-LE models, most stocks do not exhibit a significant difference in accuracy.

For the subsequent period from 2011 to 2020, as illustrated in Figures 28 and 29, all three models—HAR-J, HAR-RSV, and HAR-LE—show no significant difference in accuracy for the majority of stocks when compared to the HAR model. This contrast suggests a convergence in model performances over time. Notably, when discrepancies in accuracy do arise between these three models and the HAR benchmark, the Diebold-Mariano statistics are predominantly positive. This outcome indicates that, for specific stocks, the models HAR-J, HAR-RSV, and HAR-LE tend to outperform the HAR model, albeit this is not a universal trend across all stocks and trading periods.

While the HAR-J, HAR-RSV, and HAR-LE models, which account for volatility asymmetry, exhibit lower RMSE and MAE values than the HAR model, these improvements in raw loss measures do not consistently translate into significant accuracy enhancements, as assessed by the Diebold-Mariano test. This overall analysis suggests that introducing volatility asymmetry into the models does not uniformly enhance predictive accuracy in individual firms.

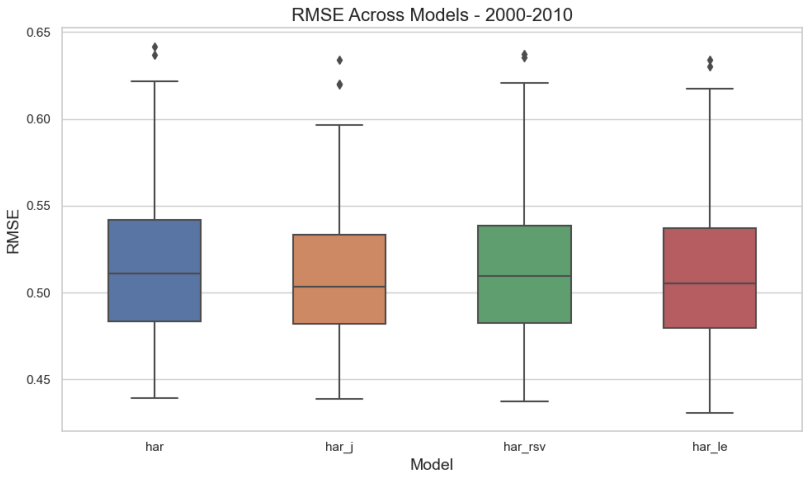

Figure 18: RMSE Across Models - 2000-2010

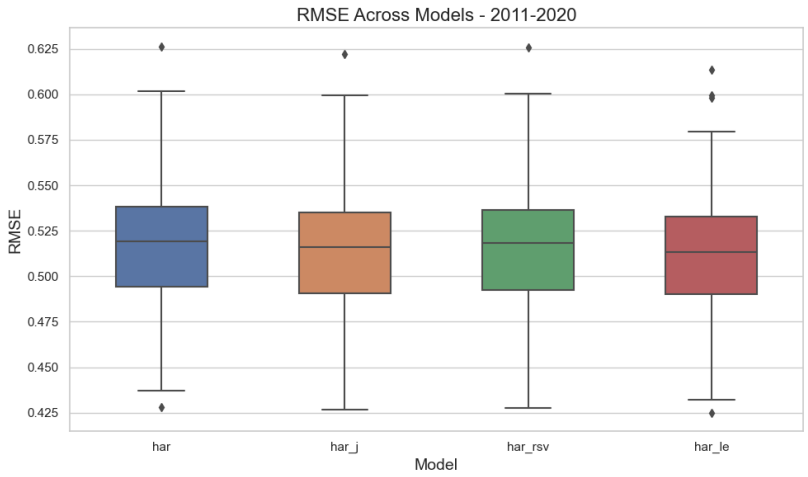

Figure 19: RMSE Across Models - 2011-2020

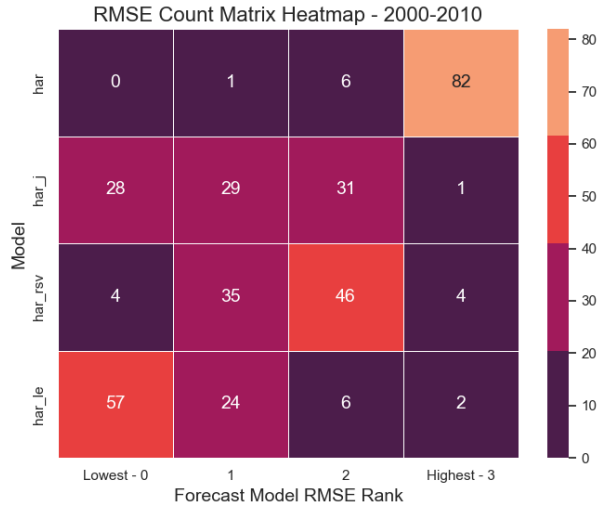

Figure 20: RMSE Count Matrix Heatmap - 2000-2010

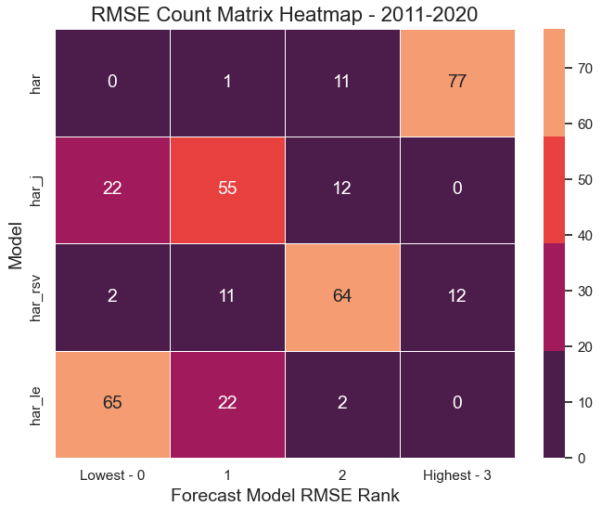

Figure 21: RMSE Count Matrix Heatmap - 2011-2020

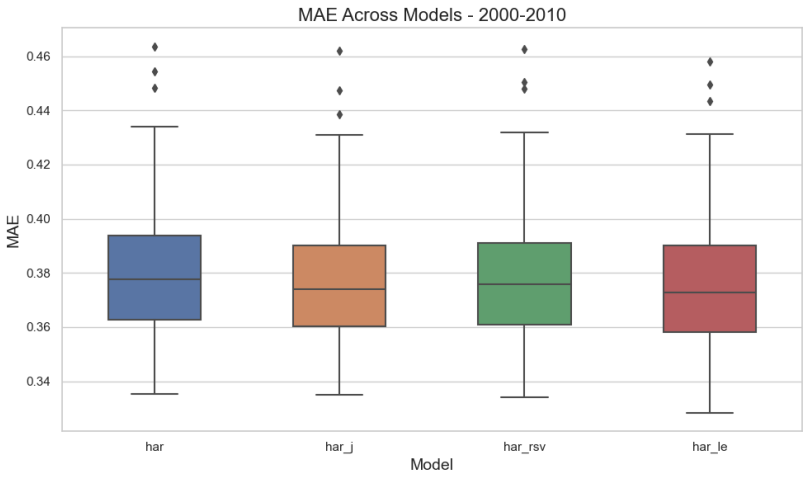

Figure 22: MAE Across Models - 2000-2010

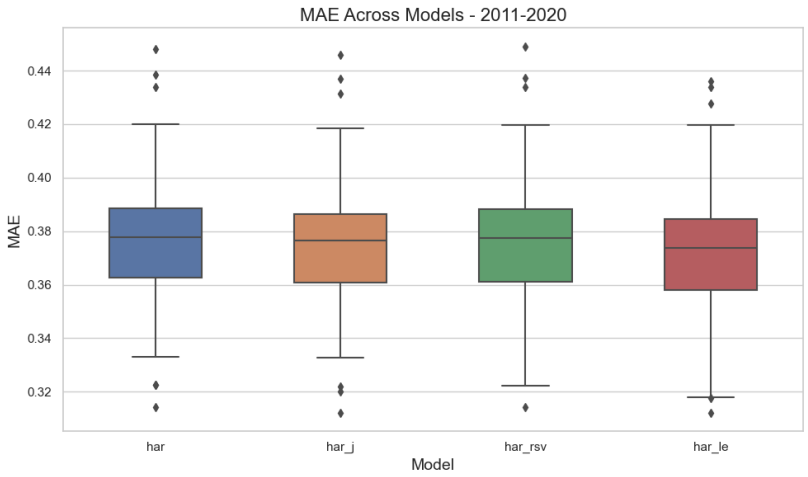

Figure 23: MAE Across Models - 2011-2020

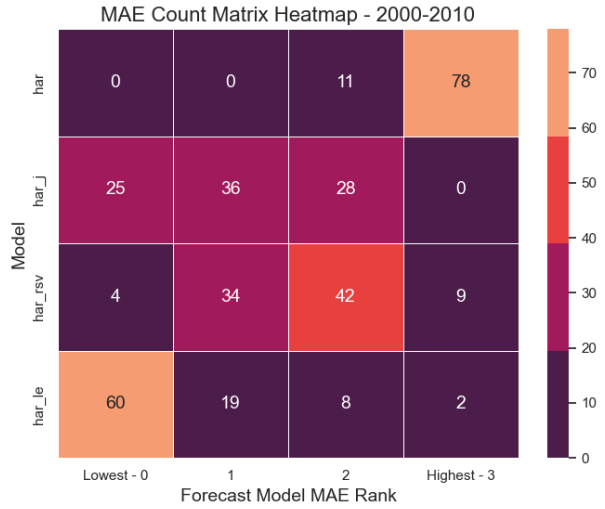

Figure 24: MAE Count Matrix Heatmap - 2000-2010

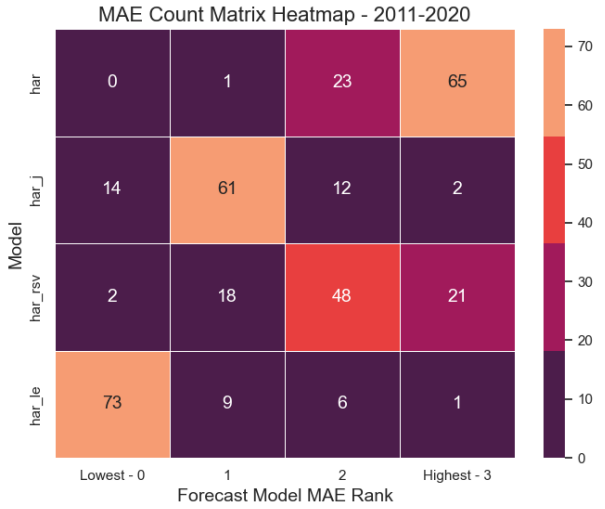

Figure 25: MAE Count Matrix Heatmap - 2011-2020

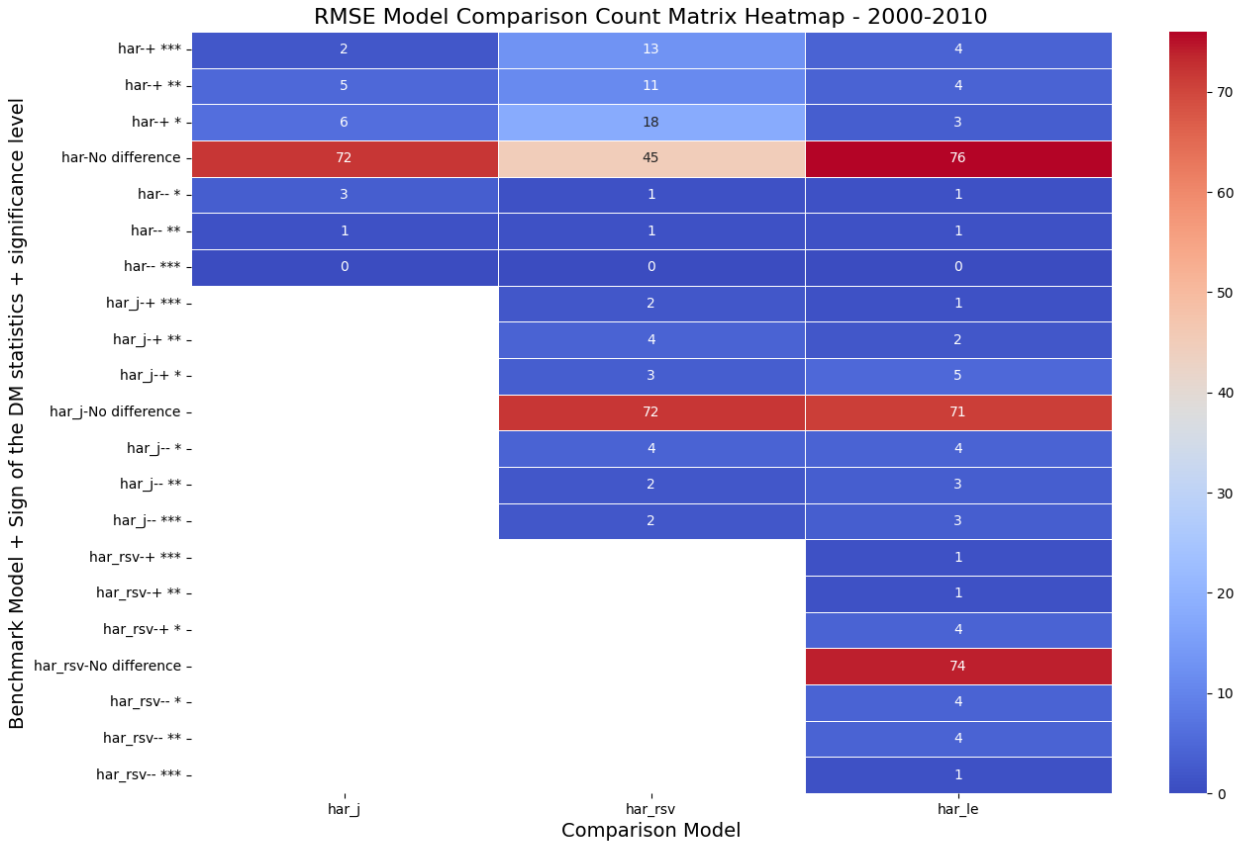

Figure 26: RMSE Diebold Mariano Model Comparison Matrix Heatmap - 2000-2010

Note: In Figure 26, the heatmap provides a quantitative comparison of forecast errors between various forecasting models, based on the Diebold-Mariano (DM) statistics across 89 stocks. The y-axis categorizes the counts into seven categories reflecting the comparative accuracy of a benchmark model against others, according to the predictive direction (better or worse) and the level of statistical significance. The x-axis lists the comparison models. Each cell shows the number of stocks where the benchmark model's forecast error, evaluated by RMSE and MAE metrics, was significantly higher (positive DM) or lower (negative DM) compared to the competing model, or where no significant difference was observed ('No difference'). The counts are further distinguished by statistical significance levels, represented by asterisks: one asterisk for 10% (*), two asterisks for 5% (**), and three asterisks for 1% (***). For instance, within the heatmap, it is noted that the HAR-RSV model exhibits significantly better forecast accuracy than the HAR benchmark model for 13 out of the 89 stocks at the 1% significance level (***).

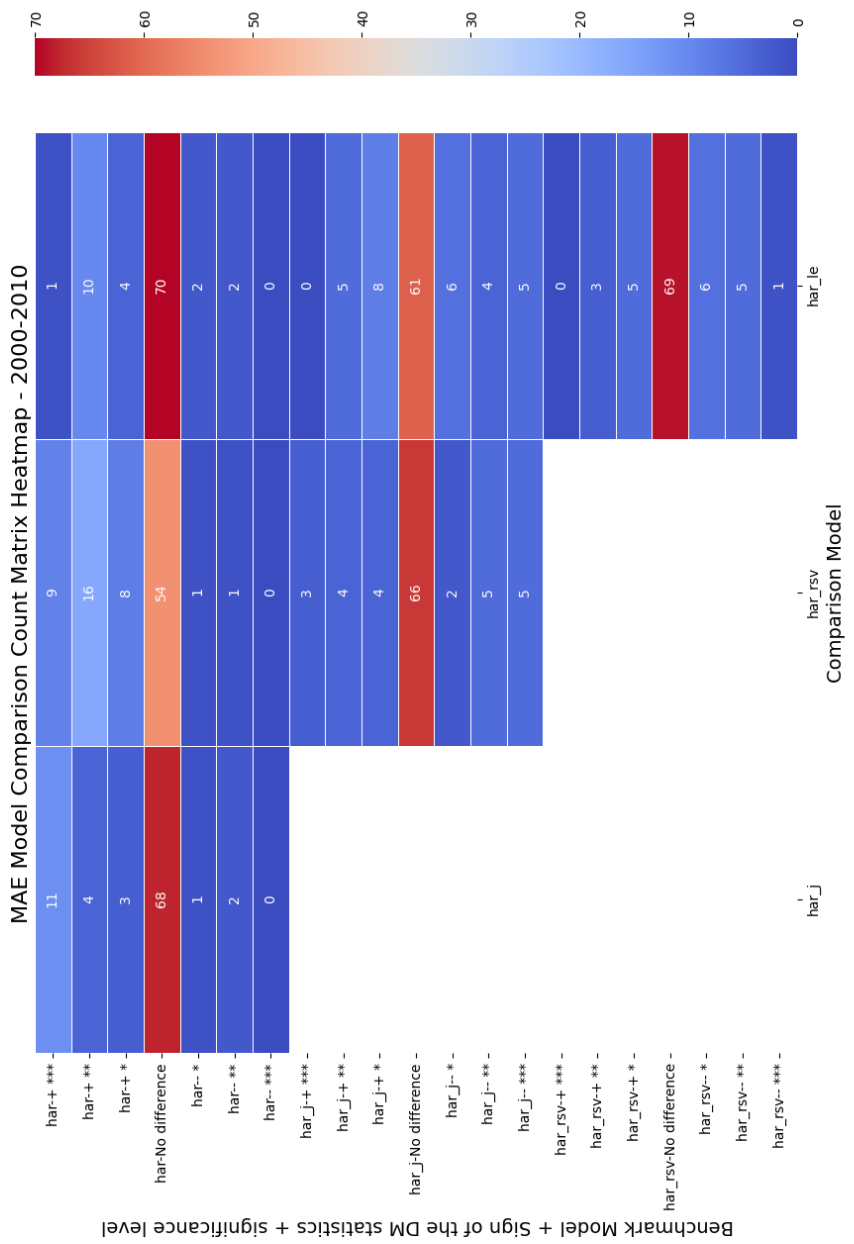

Figure 27: MAE Diebold Mariano Model Comparison Matrix Heatmap - 2000-2010

Note: See notes in Figure 26.

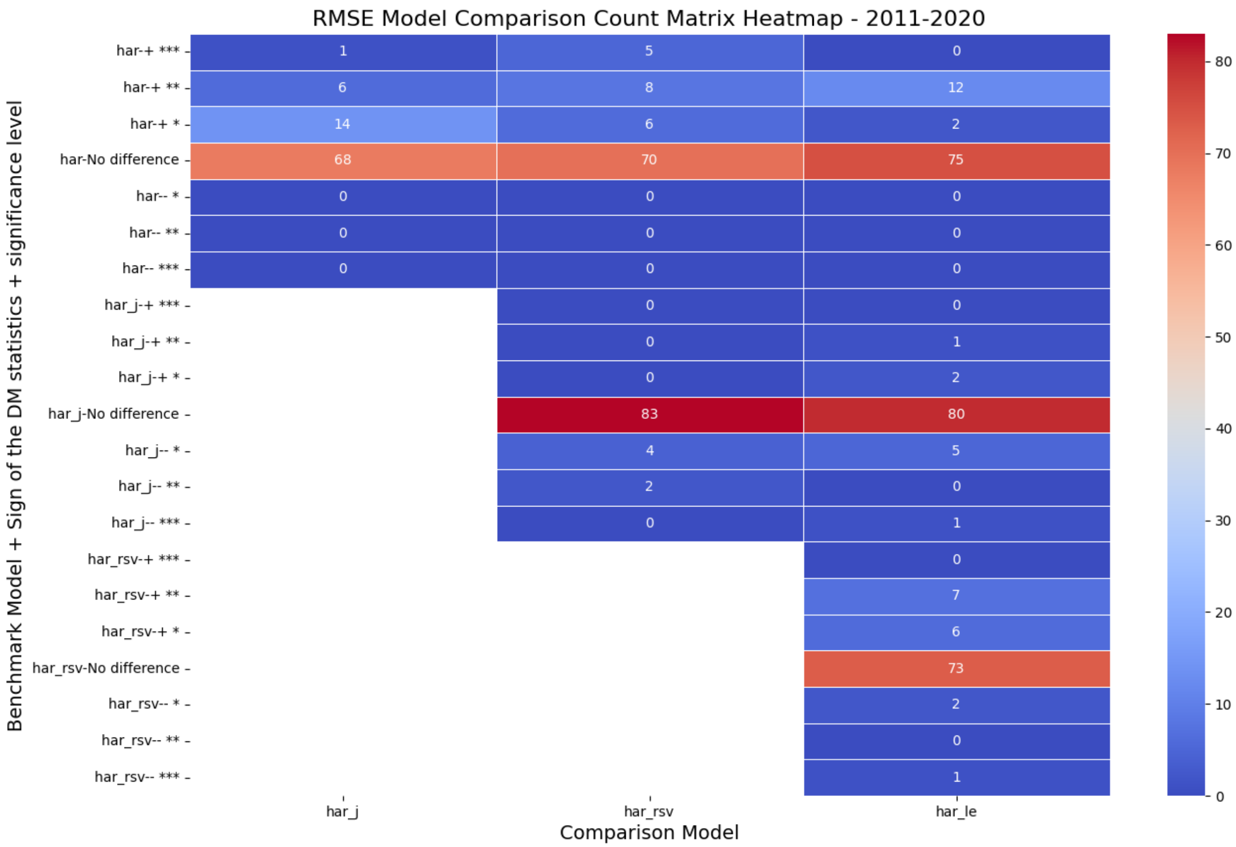

Figure 28: RMSE Diebold Mariano Model Comparison Matrix Heatmap - 2011-2020

Note: See notes in Figure 26.

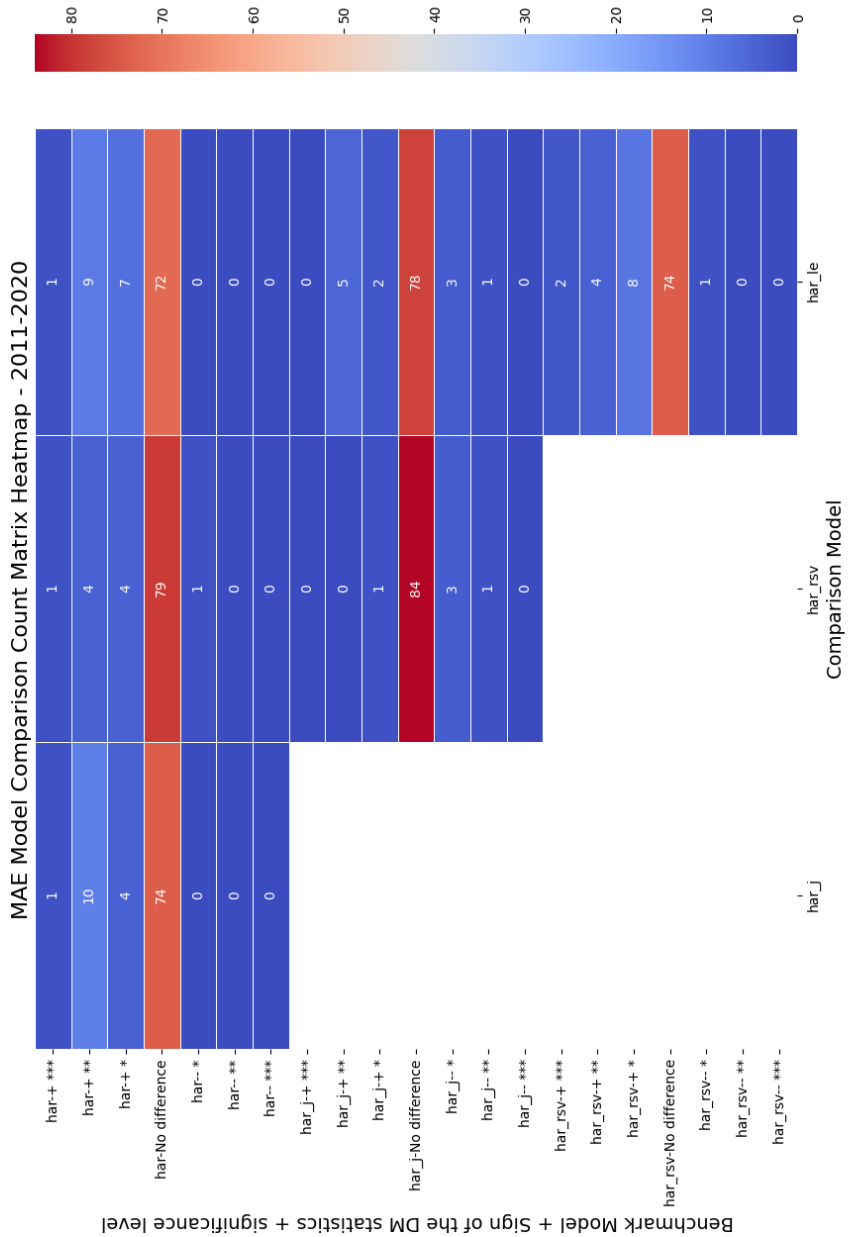

Figure 29: MAE Diebold Mariano Model Comparison Matrix Heatmap - 2011-2020

Note: See notes in Figure 26.

6 Conclusion

This thesis enhances the evaluation of the Heterogeneous AutoRegressive (HAR) model by integrating extensions such as jumps (HAR-J), realized semivariance (HAR-RSV), and leverage effects (HAR-LE). These extensions aim to improve the realized variance forecasting accuracy for the S&P 500 Index and individual stocks. An innovative aspect of this study is the division of the analysis into two distinct trading periods: 2000-2010 and 2011-2020, which provides a deeper insight into the temporal dynamics of model performance.

Our findings show that among these, the HAR-LE model provides superior forecasting performance for the S&P 500 Index, both in-sample and out-of-sample. While the out-of-sample enhancements are suggested by the raw RMSE and MAE loss functions, indicating potentially better realized variance forecasting accuracy for the HAR-J, HAR-RSV, and HAR-LE models compared to the HAR model, these improvements do not hold statistical significance at the individual stock level. This conclusion is supported by the Diebold-Mariano test, which fails to confirm superior accuracy of the extended models over the benchmark.

Our analysis confirms that incorporating elements that account for volatility asymmetries, particularly leverage effects, may significantly refines the model's predictive accuracy for the S&P 500 Index. These results not only emphasize the importance of selecting appropriate models for volatility forecasting but also highlight the benefits of adjusting for asymmetry to more effectively capture the dynamics of market volatility. Furthermore, this thesis aligns with prior studies, such as those by Andersen et al. (2007), Patton and Sheppard (2015), and Lyócsa and Todorova (2020), reinforcing the predictive strength of extended HAR models for the S&P 500 Index, while also acknowledging the mixed outcomes compared to these studies the individual stock level.

Building on these insights, there are several promising avenues for further research. One potential direction is to evaluate whether the HAR-LE model's forecasting accuracy for the S&P 500 Index during the 2000-2010 trading period surpasses that of the standard HAR model. Extending this analysis to other financial market indices, such as the Nikkei 225, and comparing results across different trading periods could also test the HAR-LE model's generalizability for market indices. Additionally, exploring the forecasting capabilities of the HAR model and its extensions over longer horizons, such as 5-day and 22-day forecasts, could provide further insights. Finally, at the individual stock level, incorporating exogenous variables that account for market sentiment and attention could further enhance the model's ability to capture volatility asymmetries, offering a comparative analysis to existing metrics

like jumps, realized semivariance, and leverage effects (Audrino et al., 2020).

References

Ampadu, S., Mensah, E.T., Aidoo, E.N., Boateng, A., Maposa, D., 2024. A comparative study of error distributions in the GARCH model through a Monte Carlo simulation approach. Scientific African 23, e01988. doi:10.1016/j.sciaf.2023.e01988.

Andersen, T.G., Bollerslev, T., 1997. Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns. Journal of Finance 52, 975–1005. doi:10.1111/j.1540-6261.1997.tb02722.x.

Andersen, T.G., Bollerslev, T., 1998. Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. International Economic Review 39, 885–905. URL: https://www.jstor.org/stable/2527343?origin=crossref, doi:10.2307/2527343.

Andersen, T.G., Bollerslev, T., Diebold, F.X., 2007. Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. Review of Economics and Statistics 89, 701–720. URL: https://direct.mit.edu/rest/article/89/4/701-720/57715, doi:10.1162/rest.89.4.701.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H., 2001. The distribution of realized stock return volatility. Journal of Financial Economics 61, 43–76. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304405X01000551, doi:10.1016/S0304-405X(01)00055-1.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 1999. (Understanding, optimizing, using and forecasting) realized volatility and correlation URL: https://archive.nyu.edu/bitstream/2451/27128/2/wpa99061.pdf.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2000. Exchange Rate Returns Standardized by Realized Volatility are (Nearly) Gaussian. URL: http://www.nber.org/papers/w7488.pdf, doi:10.17578/4-3/4-2.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2003. Modeling and Forecasting Realized Volatility. Econometrica 71, 579–625. URL: http://doi.wiley.com/10.1111/1468-0262.00418, doi:10.1111/1468-0262.00418.

Audrino, F., Sigrist, F., Ballinari, D., 2020. The impact of sentiment and attention measures on stock market volatility. International Journal of Forecasting 36, 334–357. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207019301645, doi:10.1016/j.ijforecast.2019.05.010.

Barndorff-Nielsen, O.E., Kinnebrock, S., Shephard, N., 2008. Measuring Downside Risk - Realised Semivariance. CREATES Research Paper No. 2008-4. URL: http://www.ssrn.com/abstract=1262194, doi:10.2139/ssrn.1262194.

Barndorff-Nielsen, O.E., Shephard, N., 2002. Econometric Analysis of Realized Volatility and its Use in Estimating Stochastic Volatility Models. Journal of the Royal Statistical Society Series B: Statistical Methodology 64, 253–280. URL: https://academic.oup.com/jrsssb/article/64/2/253/7098420, doi:10.1111/1467-9868.00336.

Barndorff-Nielsen, O.E., Shephard, N., 2006. Power Variation and Time Change. Theory of Probability & Its Applications 50, 1–15. URL: https://doi.org/10.1137/S0040585X97981482, doi:10.1137/S0040585X97981482.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2009. Realized kernels in practice: trades and quotes. Econometrics Journal 12, C1–C32. URL: https://academic.oup.com/ectj/article/12/3/C1/5061260, doi:10.1111/j.1368-423X.2008.00275.x.

Barndorff-Nielsen, O.E., Shephard, N., 2002. Estimating quadratic variation using realized variance. Journal of Applied Econometrics 17, 457–477. URL: https://onlinelibrary.wiley.com/doi/10.1002/jae.691, doi:10.1002/jae.691.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327. URL: https://linkinghub.elsevier.com/retrieve/pii/0304407686900631, doi:10.1016/0304-4076(86)90063-1.

Brownlees, C., Gallo, G., 2006. Financial econometric analysis at ultra-high frequency: Data handling concerns. Computational Statistics & Data Analysis 51, 2232–2245. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167947306003458, doi:10.1016/j.csda.2006.09.030.

Buncic, D., Gisler, K.I., 2017. The role of jumps and leverage in forecasting volatility in international equity markets. Journal of International Money and Finance 79, 1–19. URL: http://dx.doi.org/10.1016/j.jimonfin.2017.09.001, doi:10.1016/j.jimonfin.2017.09.001.

Clements, A., Preve, D.P., 2021. A Practical Guide to harnessing the HAR volatility model. Journal of Banking and Finance 133, 106285. URL: https://doi.org/10.1016/j.jbankfin.2021.106285, doi:10.1016/j.jbankfin.2021.106285.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. Jour-

nal of Financial Econometrics 7, 174–196. URL: https://academic.oup.com/jfec/article-lookup/doi/10.1093/jjfinec/nbp001, doi:10.1093/jjfinec/nbp001.

Corsi, F., Audrino, F., Renò, R., 2012. HAR Modeling for Realized Volatility Forecasting, in: Handbook of Volatility Models and Their Applications. Wiley, pp. 363–382. URL: https://onlinelibrary.wiley.com/doi/10.1002/9781118272039.ch15, doi:10.1002/9781118272039.ch15.

Corsi, F., Mittnik, S., Pigorsch, C., Pigorsch, U., 2008. The Volatility of Realized Volatility. Econometric Reviews 27, 46–78. URL: http://www.tandfonline.com/doi/abs/10.1080/07474930701853616, doi:10.1080/07474930701853616.

Corsi, F., Renò, R., 2009. Volatility determinants: Heterogeneity, leverage, and jumps. Technical Report. URL: https://creates.au.dk/fileadmin/site_files/filer_oekonomi/subsites/creates/Seminar_Papers/2009/FulvioCorsi_260209.pdf.

Corsi, F., Renò, R., 2012. Discrete-Time Volatility Forecasting With Persistent Leverage Effect and the Link With Continuous-Time Volatility Modeling. Journal of Business & Economic Statistics 30, 368–380. URL: http://www.tandfonline.com/doi/abs/10.1080/07350015.2012.663261, doi:10.1080/07350015.2012.663261.

Dacorogna, M.M., Müller, U.A., Pictet, O.V., Olsen, R.B., 1998. Modelling Short-Term Volatility with GARCH and HARCH Models, in: Nonlinear Modelling of High Frequency Financial Time Series, pp. 161–176. URL: http://www.ssrn.com/abstract=36960, doi:10.2139/ssrn.36960.

Diebold, F.X., Mariano, R.S., 2002. Comparing predictive accuracy. Journal of Business & Economic Statistics 20, 134–144. URL: http://www.jstor.org/stable/1392155.

Djupsjöbacka, D., 2010. Implications of market microstructure for realized variance measurement. European Journal of Finance 16, 27–43. URL: http://www.tandfonline.com/doi/abs/10.1080/13518470902853376, doi:10.1080/13518470902853376.

Engle, R.F., 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica 50, 987. URL: https://www.jstor.org/stable/1912773?origin=crossref, doi:10.2307/1912773.

Eriksson, A., Preve, D.P.A., Yu, J., 2019. Forecasting Realized Volatility Using a Nonnegative Semiparametric Model. Journal of Risk and Financial Management 12, 139. URL: https://www.mdpi.com/1911-8074/12/3/139, doi:10.3390/jrfm12030139.

Gauthier, G., 2020. Realized Variance. HEC Montréal - Méthodes d'apprentissage appliquées aux données financières - MATH80631.A2020 .

Hansen, P.R., Lunde, A., 2004. An Unbiased Measure of Realized Variance URL: http://www.ssrn.com/abstract=524602, doi:10.2139/ssrn.524602.

Hansen, P.R., Lunde, A., 2012a. Forecasting Volatility Using High-Frequency Data, in: The Oxford Handbook of Economic Forecasting. Oxford University Press, pp. 525–556. URL: https://academic.oup.com/edited-volume/28323/chapter/215083228, doi:10.1093/oxfordhb/9780195398649.013.0020.

Hansen, P.R., Lunde, A., 2012b. Realized variance and market microstructure noise. Journal of Business & Economic Statistics 24, 127–161. URL: http://www.tandfonline.com/doi/abs/10.1198/073500106000000071, doi:10.1198/073500106000000071.

Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. International Journal of Forecasting 13, 281–291. URL: https://www.sciencedirect.com/science/article/pii/S0169207096007194, doi:https://doi.org/10.1016/S0169-2070(96)00719-4.

Hendershott, T., Moulton, P.C., 2011. Automation, speed, and stock market quality: The NYSE's Hybrid. Journal of Financial Markets 14, 568–604. URL: https://linkinghub.elsevier.com/retrieve/pii/S138641811100005X, doi:10.1016/j.finmar.2011.02.003.

Horpestad, J.B., Lyócsa, S., Molnár, P., Olsen, T.B., 2019. Asymmetric volatility in equity markets around the world. North American Journal of Economics and Finance 48, 540–554. URL: https://linkinghub.elsevier.com/retrieve/pii/S1062940817303984, doi:10.1016/j.najef.2018.07.011.

Hussain, S.M., Ahmad, N., Ahmed, S., 2023. Applications of high-frequency data in finance: A bibliometric literature review. International Review of Financial Analysis 89, 102790. URL: https://linkinghub.elsevier.com/retrieve/pii/S105752192300306X, doi:10.1016/j.irfa.2023.102790.

Liu, L.Y., Patton, A.J., Sheppard, K.K., 2012. Does Anything Beat 5-Minute RV? A Comparison of Realized Measures Across Multiple Asset Classes 187, 293–311. URL: http://www.ssrn.com/abstract=2214997, doi:10.2139/ssrn.2214997.

Lynch, P.E., Zumbach, G.O., 2003. Market heterogeneities and the causal structure of

volatility. Quantitative Finance 3, 320–331. URL: http://www.tandfonline.com/doi/abs/10.1088/1469-7688/3/4/308, doi:10.1088/1469-7688/3/4/308.

Lyócsa, S., Todorova, N., 2020. Trading and non-trading period realized market volatility: Does it matter for forecasting the volatility of US stocks? International Journal of Forecasting 36, 628–645. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169207019302250, doi:10.1016/j.ijforecast.2019.08.002.

Maki, D., Ota, Y., 2021. Impacts of asymmetry on forecasting realized volatility in Japanese stock markets. Economic Modelling 101, 105533. URL: https://linkinghub.elsevier.com/retrieve/pii/S026499932100122X, doi:10.1016/j.econmod.2021.105533.

Merton, R.C., 1980. On estimating the expected return on the market: An exploratory investigation. Technical Report. National Bureau of Economic Research. Cambridge, MA. URL: http://www.nber.org/papers/w0444.pdf, doi:10.1016/0304-405X(80)90007-0.

Müller, U., Dacorogna, M., Dav, R., Pictet, O., Olsen, R., Ward, J., 1993. Fractals and intrinsic time: A challenge to econometricians, in: XXXIXth International AEA Conference on Real Time Econometrics, Luxembourg. pp. 14–15.

Müller, U.A., Dacorogna, M.M., Davé, R.D., Olsen, R.B., Pictet, O.V., Von Weizsäcker, J.E., 1997. Volatilities of different time resolutions - Analyzing the dynamics of market components. Journal of Empirical Finance 4, 213–239. URL: https://linkinghub.elsevier.com/retrieve/pii/S0927539897000078, doi:10.1016/S0927-5398(97)00007-8.

Newey, W.K., West, K.D., 1994. Automatic Lag Selection in Covariance Matrix Estimation. Review of Economic Studies 61, 631–653. URL: http://www.jstor.org/stable/2297912, doi:10.2307/2297912.

Oomen, R.C.A., 2005. Properties of Bias-Corrected Realized Variance Under Alternative Sampling Schemes. Journal of Financial Econometrics 3, 555–577. URL: https://academic.oup.com/jfec/article-lookup/doi/10.1093/jjfinec/nbi027, doi:10.1093/jjfinec/nbi027.

Patton, A.J., Sheppard, K., 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics 97, 683–697. URL: https://direct.mit.edu/rest/article/97/3/683-697/58249, doi:10.1162/REST{_}a{_}00503.

Schwert, G.W., 2011. Stock Volatility during the Recent Financial Crisis. European Financial

Management 17, 789–805. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1468-036X.2011.00620.x, doi:10.1111/j.1468-036X.2011.00620.x.

Seda, P., 2012. Heterogeneous Autoregressive Model of the Realized Volatility : Evidence from Czech Stock Market URL: http://www.wseas.us/e-library/conferences/2012/Zlin/FAA/FAA-04.pdf.

Shephard, N., 1996. Statistical aspects of ARCH and stochastic volatility, in: Time Series Models in Econometrics, Finance and Other Fields. (edited by ed.. Chapman & Hall, London, pp. 1–67. URL: https://scholar.harvard.edu/shephard/publications/statistical-aspects-arch-and-stochastic-volatility.

Shephard, N., Sheppard, K., 2010. Realising the future: forecasting with high-frequency-based volatility (HEAVY) models. Journal of Applied Econometrics 25, 197–231. URL: https://onlinelibrary.wiley.com/doi/10.1002/jae.1158, doi:10.1002/jae.1158.

Taylor, N., 2017. Realised variance forecasting under Box-Cox transformations. International Journal of Forecasting 33, 770–785. URL: http://dx.doi.org/10.1016/j.ijforecast.2017.04.001, doi:10.1016/j.ijforecast.2017.04.001.

Taylor, S.J., 2007. Modelling Financial Time Series. World Scientific Publishing Co. Pte. Ltd. URL: https://econpapers.repec.org/RePEc:wsi:wsbook:6578.

Thanasoulas, N., 2019. HAR (d) to beat? Forecasting volatility: a comparison of the HAR model and actual volatility in the Netherlands URL: https://theses.ubn.ru.nl/handle/123456789/8385.

Zhang, L., Mykland, P.A., Aït-Sahalia, Y., 2005. A Tale of Two Time Scales. Journal of the American Statistical Association 100, 1394–1411. URL: http://www.tandfonline.com/doi/abs/10.1198/016214505000000169, doi:10.1198/016214505000000169.

Zhou, B., 1996. High-Frequency Data and Volatility in Foreign-Exchange Rates. Journal of Business & Economic Statistics 14, 45–52. URL: http://www.tandfonline.com/doi/abs/10.1080/07350015.1996.10524628, doi:10.1080/07350015.1996.10524628.

A Appendix

Table 10: Final Dataset and Company Information

SYMBOL	PERMNO	CUSIP	Company Name
AAPL	14593	03783310	APPLE INC
ABT	20482	00282410	ABBOTT LABORATORIES
ADBE	75510	00724F10	ADOBE INC
AEE	24985	02360810	AMEREN CORP
AFL	57904	00105510	AFLAC INC
AN	76282	05329W10	AUTONATION INC DEL
APD	28222	00915810	AIR PRODUCTS & CHEMICALS INC
ASH	24272	04420910	ASHLAND INC NEW
ATI	43123	01741R10	ALLEGHENY TECHNOLOGIES
AZO	76605	05333210	AUTOZONE INC
BAX	27887	07181310	BAXTER INTERNATIONAL INC
BC	10874	11704310	BRUNSWICK CORP
BEN	37584	35461310	FRANKLIN RESOURCES INC
BMY	19393	11012210	BRISTOL MYERS SQUIBB CO
CAG	56274	20588710	CONAGRA BRANDS INC
CLX	46578	18905410	CLOROX CO
CMA	25081	20034010	COMERICA INC
CNX	86799	20854P10	CONSOL ENERGY INC
COST	87055	22160K10	COSTCO WHOLESALE CORP NEW
CSX	62148	12640810	C S X CORP
CTAS	23660	17290810	CINTAS CORP
DLTR	81481	25674610	DOLLAR TREE INC
EBAY	86356	27864210	EBAY INC
FAST	11618	31190010	FASTENAL COMPANY
FE	23026	33793210	FIRSTENERGY CORP
GD	12052	36955010	GENERAL DYNAMICS CORP
GPC	46674	37246010	GENUINE PARTS CO
HAS	52978	41805610	HASBRO INC
HD	66181	43707610	HOME DEPOT INC

Table 10 continued from previous page

SYMBOL	PERMNO	CUSIP	Company Name
HIG	82775	41651510	HARTFORD FINANCIAL SVCS GRP INC
HRL	32870	44045210	HORMEL FOODS CORP
HSY	16600	42786610	HERSHEY CO
IFF	40272	45950610	INTERNATIONAL FLAVORS & FRAG INC
IGT	45277	45990210	INTERNATIONAL GAME TECHNOLOGY
INTU	78975	46120210	INTUIT INC
IP	21573	46014610	INTERNATIONAL PAPER CO
IPG	53065	46069010	INTERPUBLIC GROUP COS INC
IRM	83143	46284610	IRON MOUNTAIN INC
JBHT	42877	44565810	HUNT J B TRANSPORT SERVICES INC
JCI	42534	47836610	JOHNSON CONTROLS INC
JNPR	86979	48203R10	JUNIPER NETWORKS INC
JWN	57817	65566410	NORDSTROM INC
KBH	70092	48666K10	K B HOME
KLAC	46886	48248010	K L A CORP
KMB	17750	49436810	KIMBERLY CLARK CORP
KR	16678	50104410	KROGER COMPANY
KSS	77606	50025510	KOHLS CORP
LOW	61399	54866110	LOWES COMPANIES INC
LRCX	48486	51280710	LAM RESH CORP
LUV	58683	84474110	SOUTHWEST AIRLINES CO
MAS	34032	57459910	MASCO CORP
MBI	75175	55262C10	M B I A INC
MCHP	78987	59501710	MICROCHIP TECHNOLOGY INC
MRO	15069	56584910	MARATHON OIL CORP
MTW	51263	56357110	MANITOWOC CO INC
NEM	21207	65163910	NEWMONT CORP
NI	38762	65473P10	NISOURCE INC
NSC	64311	65584410	NORFOLK SOUTHERN CORP
NVDA	86580	67066G10	NVIDIA CORP
OMC	30681	68191910	OMNICOM GROUP INC
OXY	34833	67459910	OCCIDENTAL PETROLEUM CORP
PAYX	61621	70432610	PAYCHEX INC

Table 10 continued from previous page

SYMBOL	PERMNO	CUSIP	Company Name
PBI	24459	72447910	PITNEY BOWES INC
PCAR	60506	69371810	PACCAR INC
PCG	13688	69331C10	P G & E CORP
PDCO	78034	70339510	PATTERSON COMPANIES INC
PEP	13856	71344810	PEPSICO INC
PHM	54148	74586710	PULTE GROUP INC
PPL	22517	69351T10	P P L CORP
PRGO	77182	71429010	PERRIGO CO
QCOM	77178	74752510	QUALCOMM INC
RJF	69649	75473010	RAYMOND JAMES FINANCIAL INC
ROP	77338	77669610	ROPER TECHNOLOGIES INC
SNPS	77357	87160710	SYNOPSYS INC
SNV	20053	87161C10	SYNOVUS FINANCIAL CORP
SO	18411	84258710	SOUTHERN CO
SSP	84176	81105440	E W SCRIPPS COMPANY
SWK	43350	85450210	STANLEY BLACK & DECKER INC
TEX	58318	88077910	TEREX CORP NEW
TJX	40539	87254010	T J X COMPANIES INC NEW
TMO	62092	88355610	THERMO FISHER SCIENTIFIC INC
TROW	10138	74144T10	T ROWE PRICE GROUP INC
TUP	83462	89989610	TUPPERWARE BRANDS CORP
TXT	23579	88320310	TEXTRON INC
UPS	87447	91131210	UNITED PARCEL SERVICE INC
VFC	43553	91820410	V F CORP
X	76644	91290910	UNITED STATES STEEL CORP NEW
XRAY	11600	24903010	DENTSPLY INTERNATIONAL INC NEW
YUM	85348	98849810	YUM BRANDS INC