Higashide, Takuo; Tanaka, Katsuyuki; Kinkyo, Takuji; Hamori, Shigeyuki

Article

New dataset for forecasting realized volatility: Is the Tokyo stock exchange co-location dataset helpful for expansion of the heterogeneous autoregressive model in the Japanese stock market?

Journal of Risk and Financial Management

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Higashide, Takuo; Tanaka, Katsuyuki; Kinkyo, Takuji; Hamori, Shigeyuki (2021) : New dataset for forecasting realized volatility: Is the Tokyo stock exchange co-location dataset helpful for expansion of the heterogeneous autoregressive model in the Japanese stock market?, Journal of Risk and Financial Management, ISSN 1911-8074, MDPI, Basel, Vol. 14, Iss. 5, pp. 1-18, https://doi.org/10.3390/jrfm14050215

This Version is available at:
https://hdl.handle.net/10419/239631

Journal of
Risk and Financial Management

MDPI

Article

New Dataset for Forecasting Realized Volatility: Is the Tokyo Stock Exchange Co-Location Dataset Helpful for Expansion of the Heterogeneous Autoregressive Model in the Japanese Stock Market?

Takuo Higashide [1,2], Katsuyuki Tanaka [3], Takuji Kinkyo [3] and Shigeyuki Hamori [3,*]

[1] Nissay Asset Management Department of Quantitative Investment, Tokyo 100-8219, Japan; takuo0823@gmail.com
[2] Department of Industrial and Systems Engineering, Chuo University, Tokyo 112-8551, Japan
[3] Graduate School of Economics, Kobe University, Kobe 657-8501, Japan; katsutanaka@puppy.kobe-u.ac.jp (K.T.); kinkyo@econ.kobe-u.ac.jp (T.K.)
* Correspondence: hamori@econ.kobe-u.ac.jp

Abstract: This study analyzes the importance of the Tokyo Stock Exchange Co-Location dataset (TSE Co-Location dataset) to forecast the realized volatility (RV) of Tokyo stock price index futures. The heterogeneous autoregressive (HAR) model is a popular linear regression model used to forecast RV. This study expands the HAR model using the TSE Co-Location dataset, stock full-board dataset and market volume dataset based on the random forest method, which is a popular machine learning algorithm and a nonlinear model. The TSE Co-Location dataset is a new dataset. This is the only information that shows the transaction status of high-frequency traders. In contrast, the stock full-board dataset shows the status of buying and selling dominance. The market volume dataset is used as a proxy for liquidity and is recognized as important information in finance. To the best of our knowledge, this study is the first to use the TSE co-location dataset. The experimental results show that our model yields a higher forecast out-of-sample accuracy of RV than the HAR model. Moreover, we find that the TSE Co-Location dataset has become more important in recent years, along with the increasing importance of high-frequency trading.

Keywords: realized volatility; Tokyo Stock Exchange Co-Location dataset; heterogeneous autoregressive model; random forest method; high-frequency traders

Citation: Higashide, Takuo, Katsuyuki Tanaka, Takuji Kinkyo, and Shigeyuki Hamori. 2021. New Dataset for Forecasting Realized Volatility: Is the Tokyo Stock Exchange Co-Location Dataset Helpful for Expansion of the Heterogeneous Autoregressive Model in the Japanese Stock Market? *Journal of Risk and Financial Management* 14: 215. https://doi.org/10.3390/jrfm14050215

Academic Editors: Robert Brooks and Thanasis Stengos

Received: 18 March 2021
Accepted: 5 May 2021
Published: 10 May 2021

1. Introduction

Forecasting volatility is important for financial risk management. Volatility is considered a daily varying random variable that represents the uncertainty of returns on assets. Thus, we need a more accurate volatility forecast for appropriate risk management. There are many previous studies of time-series modeling for volatility forecasting (Engle 1982; Taylor 1982; Bollerslev 1986; Nelson 1991; Glosten et al. 1993; Ding et al. 1993; Baillie et al. 1996; Harvey 1998).

Volatility is an unobservable variable, unlike returns. Andersen and Bollerslev (1998) propose using realized volatility (RV) as a proxy variable for true volatility. This is because there is a theoretical background that RV converges in probability to true volatility when the logarithmic price of assets is a semi-martingale (Barndorff-Nielsen and Shephard 2002). RV is calculated as the sum of the squares of returns observed frequently during the day. For RV forecasting, various time series models have been suggested per the heterogeneous market hypothesis proposed by Müller et al. (1997) and the discovery of the RV's long-term memory characteristics by Andersen et al. (2003), such as the fractionally integrated autoregressive moving average (ARFIMA) model proposed by Andersen et al. (2001) and the heterogeneous autoregressive (HAR) model proposed by Corsi (2009). The ARFIMA

model is well known as a long-term memory process model. In contrast, the HAR model is not a long-term memory process model but well approximates the long-term memory process with a few explanatory variables, which are past daily, weekly and monthly RV, in a linear modeling framework. Baillie et al. (2019) report that RV series are quite complex and can involve both HAR components and long memory components. Watanabe (2020) remarks that the HAR model is the most commonly used model in recent years for RV time-series modeling as the HAR model can predict RV with high prediction accuracy because of few explanatory variables. Qiu et al. (2019) remarked that the HAR model has computational simplicity (e.g., ordinary least squares method) and excellent out-of-sample performance compared to ARFIMA. Various previous studies have followed Corsi (2009), expanding and generalizing the HAR model in many directions (Andersen et al. 2007; Ubukata and Watanabe 2014; Bekaert and Hoerova 2014; Bollerslev et al. 2016; Luong and Dokuchaev 2018; Qiu et al. 2019; Motegi et al. 2020; Watanabe 2020). In particular, Luong and Dokuchaev (2018) introduced a nonlinear model using the random forest method, which is a well-known machine learning method introduced by Breiman (2001). They apply the random forest method for forecasting the direction ("up" or "down") of RV in a binary classification problem framework using a technical indicator of RV.

Linton and Mahmoodzadeh (2018) report that high-frequency trading (HFT) is the predominant feature in current financial markets due to technological advances and market structure development. Iwaisako (2017) reports that HFT has become an essential function in the stock markets of developed countries since the latter half of the 2000s. According to Iwaisako (2017), there were 81 academic papers related to HFT between 2000 and 2010, but it increased to 334 from 2011 to 2016. In the Japanese stock market, as well as in other developed countries' stock markets, the influence of high-frequency traders (HFTs) is being watched. There are some previous studies to examine the relationship between HFTs and volatility (Zhang 2010; Haldane 2011; Benos and Sagade 2012; Caivano 2015; Myers and Gerig 2015; Kirilenko et al. 2017; Malceniece et al. 2019). These existing studies report that HFTs effects on volatility. In addition, HFTs can overamplify volatility and disrupt the market with system errors. Considering the situation, the Japanese Financial Services Agency introduced the high-frequency trade participants registration system in 2018 to carefully observe these influences (The Japanese Government Financial Services Agency 2018). According to a report issued by the Japanese Financial Services Agency in August 2020, 55 investors were registered as HFT participants. In addition, 54 of the 55 investors are foreign investors. This ratio may be surprising but is only natural because about 70% of the trading in the Japanese stock market is executed by overseas investors, such as hedge funds and they adopt HFT as an edgy investment strategy.

This study analyzes the importance of the Tokyo Stock Exchange Co-Location dataset (TSE Co-Location dataset) to forecast the RV of Tokyo stock price index futures. Existing studies define the HFTs to analyze the impact of the HFTs on the volatility (Zhang 2010; Haldane 2011; Benos and Sagade 2012; Caivano 2015; Myers and Gerig 2015; Kirilenko et al. 2017; Malceniece et al. 2019). However, these existing studies may have limitation in terms of generalization. Because there is no correct answer in the definition of the HFTs (Iwaisako 2017) and the definition ambiguity remains. In this study, we respond to this problem by using the TSE Co-Location dataset. The TSE Co-Location dataset is detailed information on HFT taken by the participants who trade via a server located in the TSE Co-Location area. This server only allows participants to perform HFT. Hence, the TSE Co-Location dataset is generated with no ambiguity in the definition of HFTs. This is the only dataset that can show the actual situation of HFT of stocks in Japan. Although the HFT research is becoming more important, no analysis has been performed using the TSE Co-Location dataset. To the best of our knowledge, this study is the first to use the TSE Co-Location dataset.

We propose a new framework for forecasting the RV direction ("up" or "down") of Tokyo stock price index (TOPIX) futures in Tokyo time (9:00–15:00) using the random forest method inspired by Luong and Dokuchaev (2018). Including Loung and Dokuchaev, most

of the previous studies in RV forecast use only explanatory variables related to RV directly (e.g., viewed over different time horizons of RV, technical indicator of RV). However, in our framework, we use the past viewed RV and the TSE Co-Location dataset, stock full-board dataset and market volume dataset as explanatory variables. In particular, the TSE Co-Location dataset is one of the main characteristics of our model. The TSE Co-Location dataset is a new dataset provided by the Japan Exchange Group. This is the only dataset that can determine the activity status of HFTs in the Japanese stock market. The stock full-board dataset provides information on the potential of market liquidity and the strength of demand and supply. The market volume dataset is used as a proxy for liquidity and is recognized as important financial information. By expanding the explanatory variable space by adding these three datasets, we show that our model yields a higher out-of-sample accuracy (hereafter, we simply refer to as accuracy) of the direction of RV forecast than the HAR model through experimental results.

In summary, our main contributions of this study are as follows: First, we experimentally show the importance of the TSE Co-Location dataset to forecast the RV of TOPIX. To the best of our knowledge, this study is the first to use the TSE Co-Location dataset and show its importance. Second, our proposed model provides higher forecast accuracy than the HAR model. This is beneficial to both researchers and practitioners because it allows them to make a better selection toward the financial problem in advances. Third, we found that the random forest method framework works effectively and can be superior to the linear model in the framework of RV forecast, which is in line with the previous studies that used the random forest method for building bankruptcy models of companies (Tanaka et al. 2016, 2018a, 2018b, 2019). Our study uses a sufficiently long observation period (2012 to 2019) to consider the change in market quality affected by the HFT system and participants. Our observation period contains an essential period, which was around 2015. The HFT system named "Arrowhead" was introduced in 2010 by the Japan Exchange Group. In 2015, Arrowhead was renewed to provide a better trading system that allowed the participants to trade more frequently.

The remainder of this paper is organized as follows. In Section 2, we summarily review the previous literature. In Section 3, we briefly review the overall process of our study and introduce the details of datasets, preprocessing of datasets and the random forest method. In Section 4, we provide out-of-sample experimental results of the RV forecast accuracy. Section 5 presents the discussion and the conclusion.

2. Literature Review

2.1. Literature Review of Volatility Forecasting Models

There are many previous studies of time-series modeling for volatility forecasting, such as the autoregressive conditional heteroskedasticity (ARCH) model (Engle 1982), stochastic volatility model (Taylor 1982), generalized ARCH (GARCH) model (Bollerslev 1986), Glosten–Jagnnathan-Runkle GARCH model (Glosten et al. 1993) considering the asymmetry of volatility fluctuations, exponential GARCH (EGARCH) model (Nelson 1991) and asymmetric power GARCH (Ding et al. 1993). In addition, fractionally integrated EGARCH (Baillie et al. 1996) and stochastic volatility model with fractional integrated order (Harvey 1998) are considering long-term memory. Many other forecasting models have been studied; for example, Poon and Granger (2003) comprehensively summarize a wide range of previous studies regarding forecasting models of volatility.

Most of the volatility forecasting models are expanded based on the ARCH type modeling framework or the stochastic volatility modeling framework. On the contrary, most RV forecasting modeling is expanded based on the ARFIMA modeling framework (Andersen et al. 2001) or the HAR modeling framework (Corsi 2009). Baillie et al. (2019) assess the separate roles of fractionally integrated long memory models, extended HAR models and time varying parameter HAR models. According to Baillie et al. (2019), their experimental results suggest that RV series are quite complex and can involve both HAR components and long memory components. Recently, as we noted in the previous

section, Watanabe (2020) reports that various previous studies have followed Corsi (2009), expanding and generalizing the HAR model in many directions. Because the HAR model can predict RV with high prediction accuracy (Watanabe 2020), also the HAR model has computational simplicity and excellent out-of-sample performance compared to ARFIMA (Qiu et al. 2019). Andersen et al. (2007) proposed the HAR with continuous volatility and jumps (HAR-CJ) model, which decomposes RV into continuous and jump components, respectively, in explanatory space. Bollerslev et al. (2016) introduced the HAR quarticity (HARQ) model, which can handle the time-varying coefficients of the HAR model. In terms of asymmetric modeling, Ubukata and Watanabe (2014) and Bekaert and Hoerova (2014) proposed an asymmetric HAR model. Watanabe (2020) proposed an asymmetric HAR-CJ model and an asymmetric HARQ model. Both asymmetric models are differentiated from the symmetric model by adding a return term to the explanatory variables with a dummy. Qiu et al. (2019) proposed a versatile HAR model that applies the least-squares model averaging approach to HAR-type models with signed realized semi-variance to account for model uncertainty and to allow for a more flexible lag structure. Motegi et al. (2020) propose moving average threshold HAR models as a combination of HAR and threshold autoregression. In contrast to these linear models for RV forecasting, Luong and Dokuchaev (2018) introduced a nonlinear model using the random forest method.

2.2. Literature Review of the Relationship between HFTs and Volatility

Some studies report the relationship between HFTs and volatility. Zhang (2010) examines the implication of HFT for stock price volatility and price discovery in the US capital market using the state-space model to decompose price movements into permanent and temporary components and to relate changes in both to HFT. According to Zhang (2010), HFT is positively correlated with stock price volatility. In addition, the positive correlation is stronger during periods of high market uncertainty. Zhang (2010) essentially defines HFT as all short-term trading activities by hedge funds and other institutional traders not captured in the 13F database.

Haldane (2011) reports that HFT algorithms tend to amplify cross-stock correlation in the face of a rise in volatility due to their greater use of algorithmic trend-following and arbitrage strategies.

Benos and Sagade (2012) analyze the HFTs impact on excess volatility in the UK equity market using the multivariate analysis tests for contemporaneous causal effects of spread and volatility on HFT activity. This study defines two types of volatility; “good” (when price changes reflect the arrival of new information about fundamentals) or “excessive” (when price changes do not reflect any information about fundamentals). As experimental results, HFT contributes a large amount of both “good” and “excessive” volatility. In this study, Benos and Sagada note as follows; “Although there is no precise definition of an HFT, the term is commonly used to describe firms that use computers to trade at high speeds and who also tend to end the day flat, i.e., carry small or no overnight positions. This paper uses a sample from a data set of transaction reports, maintained by the Financial Services Authority.”

Caivano (2015) studies the impact of HFT on stock price volatility over the period 2011–2013 for a sample of 5 blue chips traded on Borsa Italiana. In order to analyze the impact of HFT on volatility, this study implements a panel two-stage instrumental variables fixed effect estimation. Results show that an exogenous increase of HFT activity causes a statistically and economically significant increase in volatility. In detail, if HFT activity increases by ten percentage points, the annualized intraday volatility increases by an amount between 4 and 6 percentage points depending on the specification used.

Myers and Gerig (2015) developed an agent-based simulation and showed that HFT theoretically reduces volatility. Kirilenko et al. (2017) report that HFTs generally follow the trend four- seconds after an event, then reverse the trade after 10-s and this behavior may lead to an increase in volatility. This study uses audit trail account-level transaction data in the E-mini S&P 500 stock index futures. Malceniece et al. (2019) report that

HFT is associated with increased volatility through the experimental results in European equity markets.

3. Materials and Methods

Figure 1 shows the entire process of the experimental procedure. First, we processed the raw datasets used in our research: NEEDS Tick Data File, which contains a stock full-board dataset and market volume dataset and the TSE Co-Location dataset, which contains HFT information. In preprocessing 1, the NEEDS Tick Data File, which is the high-frequency dataset, is thinned to reduce market micro-structured noise; then, we calculate RV and organize the stock full-board dataset. For the TSE Co-Location dataset, we calculate the ratio of all transactions traded from the TSE Co-Location server. By doing this preprocessing, we can determine the approximate effect of HFTs in the entire Japanese stock market. In preprocessing 2, we perform standardization so that different sizes of data can be analyzed in the same explanatory variable space. In addition, considering investors with various investment horizons, following the HAR model, we calculate the previous day data and the average of the past data (weekly and monthly, which correspond to 5 and 22 working days, respectively) for each variable to put in the explanatory space to predict RV.

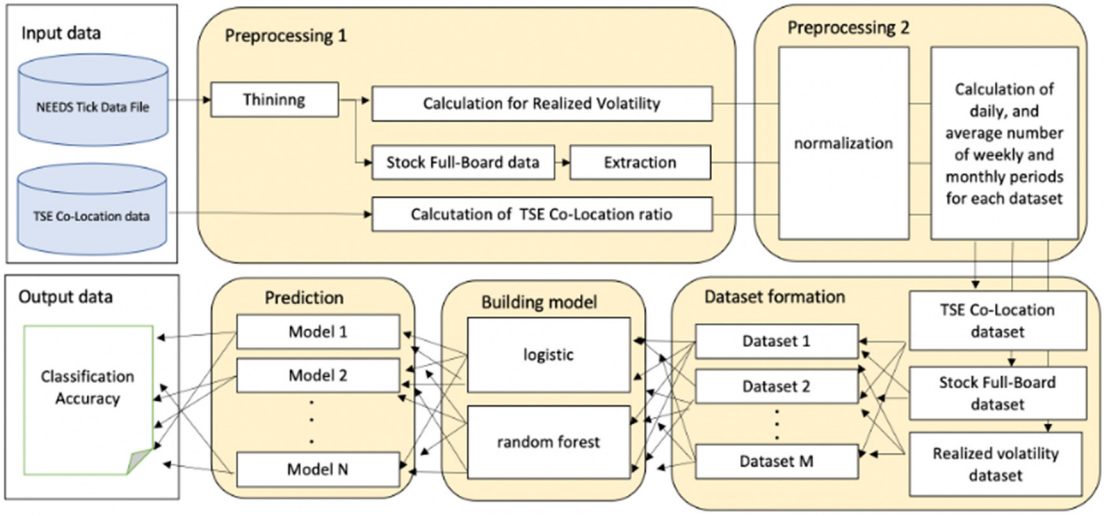

Figure 1. Overall process.

In the dataset formation process, we created several combinations of explanatory variable datasets to analyze the contribution of each variable to the improvement of prediction accuracy.

Finally, we built models for each dataset using the logistic and random forest methods to compare their RV forecast accuracy. We provide the prediction accuracy from different methods and tasks. Below, we describe the details of each preprocessing and dataset.

3.1. NEEDS Tick Data File Preprocessing

We use the NEEDS Tick Data File provided by NIKKEI Media Marketing (NIKKEI Media Marketing NEEDS Tick Data n.d.) for RV calculation (Section 3.1.1) and stock full-board dataset preprocessing (Section 3.1.2). Before calculation and preprocessing, we thin out every 5 min according to previous studies (Ubukata and Watanabe 2014). Most previous researchers reported that the smaller the interval, the larger the market microstructure noise that may be contained during the RV calculation. We extract the following information: traded price, traded volume and stock full-board dataset, which is composed of the 1st best quote to the 10th best quote quantity and price on both the bid and offer sides. In the morning session, the data points we extracted were 09:01, 09:05, . . . , 11:25. In the afternoon session, 12:31, 12:35, . . . , 14:55. Note that there is only a morning

session on both the grand opening and closing. Therefore, we extracted only the morning session on these two days.

3.1.1. Realized Volatility Calculation

Given return data r_t, $r_{t+1/n}, \ldots, r_{t+(n-1)/n}$ of intraday on t, where n is the sample size within a day, RV is calculated by

$$RV_t^{(d)} = \alpha \sum_{i=0}^{n-1} r_{t+i/n}^2 \tag{1}$$

where

$$\alpha = \sum_{t=1}^{T} \left(R_t - \overline{R}\right)^2 / \sum_{t=1}^{T} RV_t \tag{2}$$

Here, the subscript t indexes the day, while T indexes the endpoint within the observation period. α indexes the evening time-adjustment coefficient. The superscript (d) in Equation (1) indexes daily. We follow Watanabe (2020) to calculate Equation (2), as proposed by Hansen and Lunde (2005). Note that we calculate the return for RV based on the trade price. If there are no transactions, we use the previously traded price.

3.1.2. Stock Full-Board Dataset Preprocessing

The stock full-board dataset is an important piece of information that provides valuable market conditions, such as the potential of market liquidity and the strength of demand and supply. These market conditions can affect volatility, which represents price fluctuation. In addition, a large bias in supply and demand suggests that the market is more likely to move in one direction (that is, a sign of a trend forming) and volatility may increase. This is well known from a practical point of view. Some previous studies use a stock full-board dataset to forecast returns or volatility in the Japanese market (Toriumi et al. 2012).

When thinning out every 5 min, we follow the procedure below to extract information as explanatory variables from the stock full-board dataset. For each five min-period, we extract the 1st best quote to the 10th best quote quantity when either the price of the 1st best quote changes or is traded. Then, we take the summation of the 1st best quote to the 10th best quote quantity, standardized by the traded quantity. If there are no transactions, we use the previously traded price. Let Bid_t, $Offer_t$ at the datapoint of t be the summation on both the bid side and offer side standardized quantity above. Then, we calculate using the following Equation: Suppose given data Bid_t, $Bid_{t+1/n}$, …, $Bid_{t+(n-1)/n}$ and $Offer_t$, $Offer_{t+1/n}$, …, $Offer_{t+(n-1)/n}$.

$$\text{Cum\_Plus}_t = \sum_{i=0}^{n-1} Bid_{t+i/n} + Offer_{t+i/n}, \tag{3}$$

$$\text{Cum\_Minus}_t = \sum_{i=0}^{n-1} Offer_{t+i/n} - Bid_{t+i/n} \tag{4}$$

Equation (3) describes the liquidity and Equation (4) describes the demand and supply of the market, respectively.

3.2. TSE Co-Location Dataset Preprocessing

The TSE Co-Location dataset (Japan Exchange Group Connectivity Services 2021) is a new dataset that delivers HFTs trading information provided by the Japan Exchange Group. More specifically, it provides detailed information on HFT and is composed of order quantity, order to execution quantity and value traded quantity taken by the participants who trade via a server located in the TSE Co-Location area. This server only allows participants to perform HFT. In other words, the TSE Co-Location dataset is the aggregated information of transactions conducted through this server and this is the only dataset

that can show the actual situation of HFT of stocks in Japan. Therefore, it is clear that the co-location information is closely related to high-frequency price data. Thus, it is an important variable that generates RV. As mentioned earlier, the importance of HFTs actions has been increasing annually since 2010. Under this trend, the TSE Co-Location dataset is the only important bridge for researchers and practitioners to consider the influence of HFTs in the Japanese stock market.

In this study, we use three explanatory variables on day t. Note that there are only two ways to trade Japanese stocks on the Tokyo Stock Exchange market: via TSE Co-Location or the other. Thus, each denominator of Equations (5)–(7) is the total number taken by these two methods. In contrast to the denominator, the numerator shows only the number taken through the TSE Co-Location server.

$$Colo\_C = \frac{order\ quantity\ via\ TSE\ Co-Location\ area}{total\ order\ quantity}, \tag{5}$$

$$Colo\_Y = \frac{order\ to\ execution\ quantity\ via\ TSE\ Co-Location\ area}{total\ order\ of\ execution\ quantity}, \tag{6}$$

$$Colo\_B = \frac{value\ traded\ quantity\ via\ TSE\ Co-Location\ area}{total\ value\ traded\ quantity}. \tag{7}$$

We also use the total value traded quantity as a single explanatory variable in our model, which is the denominator of Equation (7). This variable enables practitioners to understand a market activity or liquidity. It may be called the market volume data among practitioners. In this paper, we briefly describe this variable as follows.

$$market\ volume := total\ value\ traded\ quantity. \tag{8}$$

3.3. Dataset Formation

We built five types of models: (I) HAR model, (II) HAR + Volume model, (III) HAR+ TSE Co-Location model, (IV) HAR+ Stock full board model and (V) HAR+ Volume + TSE Co-Location + Stock full board model, using the dataset mentioned in Sections 2.1 and 2.2, as shown in Table 1. As we mentioned in the overall process explanation, we prepare previous day data (which is denoted by "_daily") and two different averages of past data, which are weekly and monthly (which is denoted by "_weekly" and "_monthly," respectively), for each variable. By using these five different models, we examined how each variable contributes to the improvement of RV prediction accuracy.

Table 1. Dataset.

HAR	Volume	TSE Co-Location	Stock Full-Board
RV_daily	market volume_daily	Colo_C_daily	Cum_Plus_daily
RV_weekly	market volume_weekly	Colo_Y_daily	Cum_Minus_daily
RV_monthly	market volume_monthly	Colo_B_daily	Cum_Plus_weekly
		Colo_C_weekly	Cum_Minus_weekly
		Colo_Y_weekly	Cum_Plus_monthly
		Colo_B_weekly	Cum_Minus_monthly
		Colo_C_monthly	
		Colo_Y_monthly	
		Colo_B_monthly	

Incidentally, the HAR model is known as the linear regression model in Equation (9),

$$\log RV_{t+1} = \beta_0 + \beta_d \log RV_t^{(d)} + \beta_w \log RV_t^{(w)} + \beta_m \log RV_t^{(m)} + \varepsilon_{t+1}, \tag{9}$$

where

$$\log RV_t^{(l)} = l^{-1} \sum_{s=1}^{l} \log RV_{t-s}.$$

This Equation is a linear combination of the constant term, daily RV (which is denoted by $RV_t^{(d)}$) calculated by Equations (1) and (2), weekly RV and monthly RV. Indicating the aggregation period as a superscript, the notation for the weekly RV is $RV_t^{(w)}$, while the monthly RV is denoted by $RV_t^{(m)}$. For instance, the weekly RV at time t is given by the average

$$RV_t^{(w)} = \frac{1}{5}\left(RV_t^{(d)} + RV_{t-1d}^{(d)} + \cdots + RV_{t-4d}^{(d)}\right).$$

The HAR model is not a long-term memory process but has three different types of autoregressive terms to approximate long-term memory processes.

As shown in Table 1, we prepare different types of autoregressive terms for the explanatory variables. Note that these explanatory variables are the rate of change. Using these datasets, we build five different types of models and forecast the RV direction ("up" or "down"). For this, we define the RV direction as follows:

$$\delta_t = \begin{cases} 1\ (up) & if\ \frac{RV_t}{RV_{t-1}} > 1 \\ 0\ (down) & if\ \frac{RV_t}{RV_{t-1}} < 1 \end{cases}.$$

If $RV_t / RV_{-1} = 1$, this case is omitted from the training dataset. In this study, there is no data point for this case.

3.4. Methods

Random forest is a popular machine learning method used for classification and regression tasks with high-dimensional data (Breiman 2001). Random forests are applied in various areas, including computer vision, finance and bioinformatics, because they provide strong classification and regression performance. Random forest is called ensemble learning in the field of machine learning because random forest combines and aggregates several predictions outputted by several randomized decision trees. Each decision tree corresponds to a weak discriminator in ensemble learning. Random forests are structured through an ensemble of d decision trees with the following algorithm:

1. Create subsets of training data with random sampling by bootstrap.
2. Train a decision tree for each subset of training data.
3. Choose the best split of a variable from only the randomly selected m variables at each node of the tree and derive the split function.
4. Repeat steps 1, 2 and 3 to produce d decision trees.
5. For test data, make predictions by voting or by averaging the most popular class among all of the output from the d decision trees.

The Gini index proposed by Economist Gini is a popular evaluation criterion for constructing decision trees (Breiman et al. 1984), where the Gini index is used to measure the impurity of each node for the best split. The criteria of the best split are determined to maximize the decline rate of impurities at each node. The Gini index is an essential criterion for selecting the optimal splitting variable and the corresponding threshold value at each node. Suppose M_n is the number of pieces of information reaching node n and M_n^i is the number of data points belonging to class C_i. The Gini index, GI_n, of node n is

$$GI_n = 1 - \sum_{i=1}^{k} \left(p_n^i\right)^2, where\ p_n^i = \frac{M_n^i}{M_n}.$$

A higher Gini index value for node n represents an impurity. Hence, a decreasing Gini index is an important criterion for node splitting.

While the random forest method is not state-of-the-art, such as deep learning techniques, we choose this method as it has several preferable features. First, random forests provide higher classification accuracy because they integrate a large number of decision trees. Second, random forests are robust to over-fitting because of the bootstrap sampling of data and random sampling of variables to build each decision tree; hence, the correlation between decision trees is low. As a result, the effect of overfitting is extremely small and the generalization ability is enhanced. Third, random forests can handle large datasets without needing too much calculation time because they enable researchers to train multiple trees efficiently in parallel. Moreover, unlike deep learning, the number of hyper-parameters is small and researchers do not need to puzzle over the hyper-parameter settings; researchers only need to choose the number of decision trees to build a model. Finally, random forests can be used to rank the importance of variables, which helps researchers identify the influential variables in the model. Therefore, researchers can manage the model efficiently and explain the contents of the model to stakeholders.

4. Experimental Results

In this section, we present the experimental results. The TSE Co-Location dataset may not be familiar to both practitioners and researchers; hence, we show the time series chart of the TSE Co-Location dataset and RV in Figure 2 and discuss its implications.

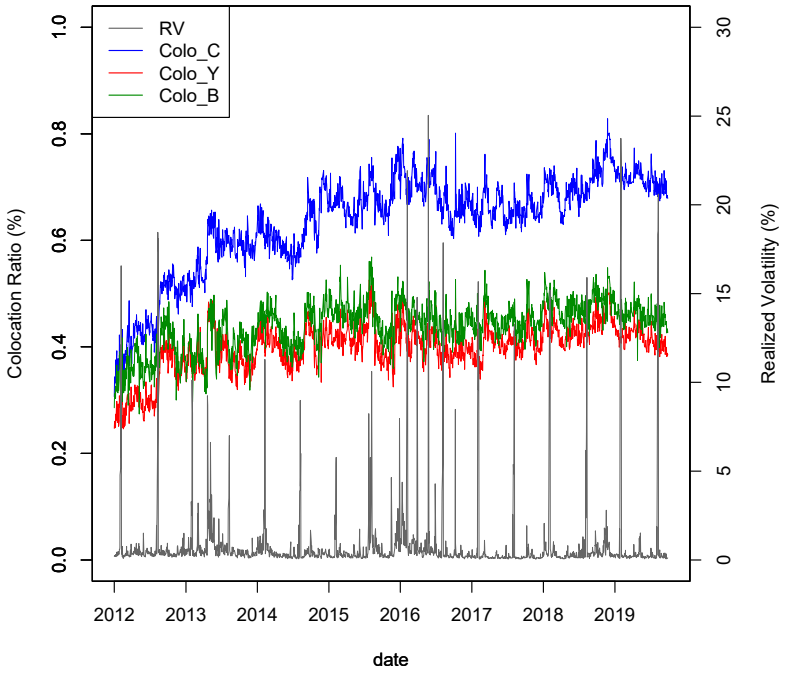

Figure 2. Time series of the TSE Co-Location dataset and RV. In this figure, RV denotes $RV_t^{(d)}$, while Colo_C, Colo_Y and Colo_B denote co-location ratios; the ratio of order quantity via the TSE Co-Location area to total order quantity (defined in Equation (5)), the ratio of execution quantity via the TSE Co-Location area to total order of execution quantity (defined in Equation (6)) and the ratio of the value traded quantity via TSE Co-Location area to total value traded quantity (defined in Equation (7)), respectively.

As can be seen from this figure, in the early 2010s, each TSE Co-Location index gradually increased because Arrowhead was introduced in 2010. This implies that the number of HFTs gradually increased during the system transition period. That is, the adjustment time to a new system varies from practitioner to practitioner. Thus, each index continues to increase for a certain period. A few years after its introduction, the Arrowhead system was revised in 2015. This revision allows trading participants to trade faster and

more frequently. As a result, Colo_C continued to rise from the middle of the 2010s to the end of the 2010s. In contrast, Colo_B and Colo_Y have been flat since 2015.

As noted in the previous section, most HFTs in the Japanese stock market are foreign investors. Thus, we are of the opinion that these three TSE Co-Location numbers refer mainly to foreign investors. Since 2010, the Abenomics policy has attracted foreign investors' interest in the Japanese stock market. It is known that foreign investors account for approximately 70% of the Japanese stock market (HFT and regular trading).

4.1. Observation Period

Our observation period is from 1 March 2012 to 31 October 2019. This period covers an essential event, that is, the renewal of Arrowhead on 24 September 2015. In addition, the beginning of our observation is not far from the implementation of the Arrowhead system. Namely, our observation period covers most of the period since HFT became possible in Japan. Thus, our examination of the accuracy of the model is highly reliable. We split the data into training data and test data with a 9:1 ratio. When building the model, we must consider data bias. If data are biased during the training period, the model is biased. Table 2 shows the sample size of "up" and "down." From Table 2, our experimental result is worthy of discussion because there is no bias in the training data.

Table 2. Sample size.

	Total Observation Period	
	Down	**Up**
training data	848	841
test data	93	94

4.2. Experimental Results and Consideration

4.2.1. RV Prediction Accuracy

In this sub-sub section, we examine the different types of models by RV forecast accuracy and evaluate the effect of each dataset on the improvement of RV forecast accuracy through various experimental patterns. To evaluate the prediction accuracy, we used the F-measure, which is the harmonic mean that summarizes the effectiveness of precision and sensitivity in a single number. It is commonly used for evaluating the accuracy of classification (Croft et al. 2010; Patterson and Gibson 2017).

$$F - measure = 2 * precision * sensitivity / (precision + sensitivity).$$

Here, precision measures the number of correct predictions divided by all instances. Sensitivity measures the number of correct predictions divided by all correct instances. Table 3 shows the prediction accuracy of the RV during the total observation period for each model.

Table 3. Prediction accuracy of RV in the total observation period.

Total Observation Period			
No	**Model**	**F-Measure**	
		Random Forest	**Logistic**
I	HAR	0.60	0.59
II	HAR + Volume	0.64	0.52
III	HAR + TSE Co-Location	0.63	0.53
IV	HAR + Stock full board	0.66	0.46
V	HAR + Volume + TSE Co-Location + Stock full board	0.68	0.46

In the explanatory variables, the space of the HAR model (denoted as NoIin Table 3; below, we unify this notation rule and others, as well), logistic method and our suggested random forest method yield almost the same prediction accuracy. However, with an increase in explanatory variables, the difference in prediction accuracy between the logistic method and the random forest method is larger. For instance, there is a 22% difference inaccuracy in the HAR + Volume + TSE Co-Location + Stock full-board model. These results are consistent with those of previous studies (Ohlson 1980; Shirata 2003; Hastie et al. 2008; Alpaydin 2014), which reported that linear models do not work where there are many explanatory variables in space.

Next, we shift to the differences between the models based on the RF method. The HAR model yielded a forecast accuracy of 60%. On the contrary, the HAR + Volume model yielded a 4% higher accuracy than the HAR model. In addition to the HAR + Volume model, the HAR + TSE Co-Location model and the HAR + stock full-board model yielded 3% and 6% higher accuracy, respectively. Moreover, the HAR + Volume + TSE Co-Location + Stock full-board model was 8% higher. From these results, it is evident that each explanatory variable helps the HAR model to improve forecast accuracy. In particular, the finding of an 8% increase in the prediction accuracy is a major contribution. However, Table 3 implies that the features of these three variables may partially overlap. Compared to the total of 13%, which is simply the sum of the effects of each explanatory variable, the HAR + Volume + TSE Co-Location + Stock full-board model yields 5% lower than 13%. Future studies could work toward elucidating the aspects in which each variable has overlapping features despite having different data generating processes.

4.2.2. Analyzing Important Variables

The order of important variables helps us gain an in-depth understand and confirm the important variables. This information is especially beneficial for practitioners because they have to explain the content of the model to customers or supervisors. The random forest method generates the importance of each variable. This visualization is one of the advantages of the random forest method. Figure 3 shows the importance variables arranged in descending order based on the Gini index in the building process of the HAR + Volume + TSE Co-Location + Stock full-board model. Thus, we know that RV_daily is the most important variable in this model. RV_monthly and RV_weekly are the second and third most important variables, respectively.

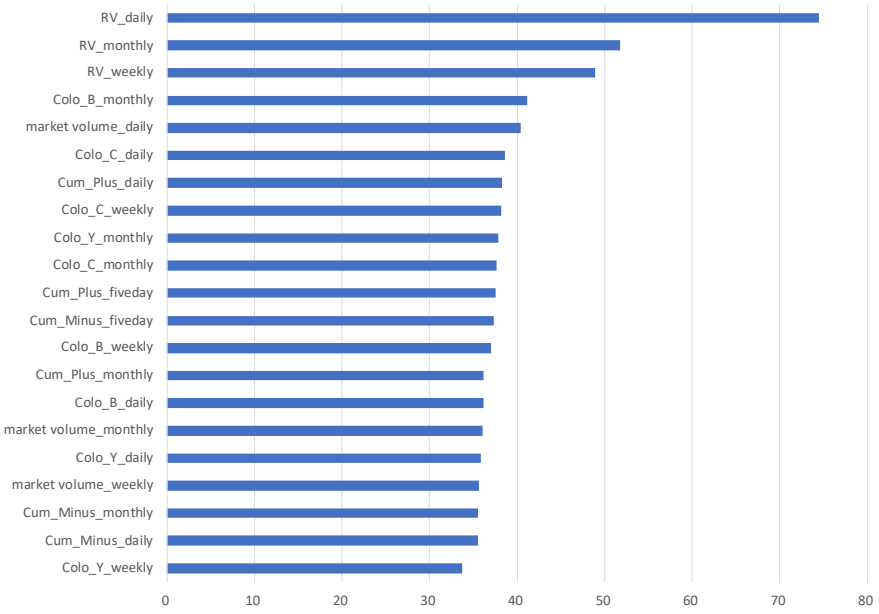

Figure 3. Important variables in the total observation period.

The top three important variables are RV's autoregressive terms in different time horizons. This result is natural because there is a clustering effect on volatility. The RV's past data are beneficial information for the forecast itself. Interestingly, five of the Top 10 important variables are the TSE Co-Location dataset. This accounts for over 70% of the Top 10 variables. We can state that the TSE Co-Location dataset is especially important for improving the prediction accuracy of the HAR + Volume + TSE Co-Location + Stock full-board model. Especially for Colo_C, all types of time horizons rank in. In contrast, for Colo_B and Colo_Y, the monthly time horizon is ranked. The results indicate that long-term trends are more important than short-term trends for these two variables. Market volume_daily and Cum_Plus_daily were also ranked in the Top 10 important variables. Both variables are related to market liquidity. The difference between the two variables is the amount actually traded or the amount that can be traded. As mentioned above, liquidity is closely related to volatility. Higher liquidity stabilizes the market environment, where prices are less likely to jump and volatility is more stable. This result is consistent with a practical point of view. Contrarily, the results of the importance of Cum_minus are rather surprising to practitioners. From a practical point of view, Cum_minus is known as an important variable in looking at the future direction of the market and indicates the strength of supply and demand between selling and buying. Our results suggest that the price direction may not be directly related to the up and down forecast of RV since the ranking of Cum_Minus_weekly, Cum_Minus_monthly and Cum_Minus_daily is not very high; they are ranked 12, 19 and 20 out of 21 variables, respectively, in the important variables.

From another perspective, we look at the important variables from the time horizon: daily, weekly and monthly. Table 4 shows the rank of the periods in descending order by the Gini index for each category. In most categories, the daily period was ranked at the top. We cannot always necessarily say that the shorter the period, the more important it is. The comparison between weekly and monthly data is more important than weekly data. From this case, it is evident that very short periods and slightly longer periods play a more important role in the model. However, it depends on the category, but the tendency is as noted above. One possible explanation is that the expiration of information, which is aggregated by these categories, is nonlinear in RV forecast in the Japanese stock market. Detailed discussions require larger and more extensive analyses, such as comparing trends among countries.

Table 4. Comparison of the importance of periods in each category.

Frequency	RV	Market Volume	Colo_C	Colo_Y	Colo_B	Cum_Plus	Cum_Minus	Average
daily	1	1	1	1	3	1	2	1.4
weekly	3	2	2	3	2	1	1	2.0
monthly	2	2	2	2	1	2	2	1.9

Note: 1, 2 and 3 denote the rank of each important variable. For example, among the RV, RV_daily is the most important variable. RV_monthly and RV_weekly are the second and third most important variables, respectively. The average is the average rank of these seven categories.

4.2.3. Examination of TSE Co-Location Dataset Importance

To examine the effect of the TSE Co-Location, we split the total observation period (1 March 2012 to 31 October 2019) into two periods to consider the effect of Arrowhead renewal in 2015: the first half period (1 March 2012 to 23 September 2015) and the second half period (24 September 2015 to 31 October 2019). For each period, we build a model in the same framework as in the previous sub-sub section to compare the differences in important variables between the first and second half periods from a time-series perspective. Figures 4 and 5 show the important variables in the first and second half periods, sorted in descending order based on each period. Considering the changes in the importance of variables, RV remains an important variable in both the first and second half periods.

However, in the categories excluding RV, the importance of the TSE Co-Location dataset increased overall and ranked higher from the first half period to the second half period. In particular, the increase in Colo_C was remarkable. Colo_C occupies the second position in the second half period, following RV. Colo_B ranks in the top three, but this category is less important in the TSE Co-Location dataset than in the first half period. This suggests that information on the order status of HFT among market participants is more valuable than that of what is bought or sold. It is interesting to recognize this trend as an increase in HFTs. In contrast, Colo_Y was not as important in both periods. In fact, remember the flow of order → execution → trading volume, when an order is filled, that number is reflected in the trading volume. From this, we think that it is possible to interpret that Colo_Y is not an important variable because it has a strong meaning between order and trading volume.

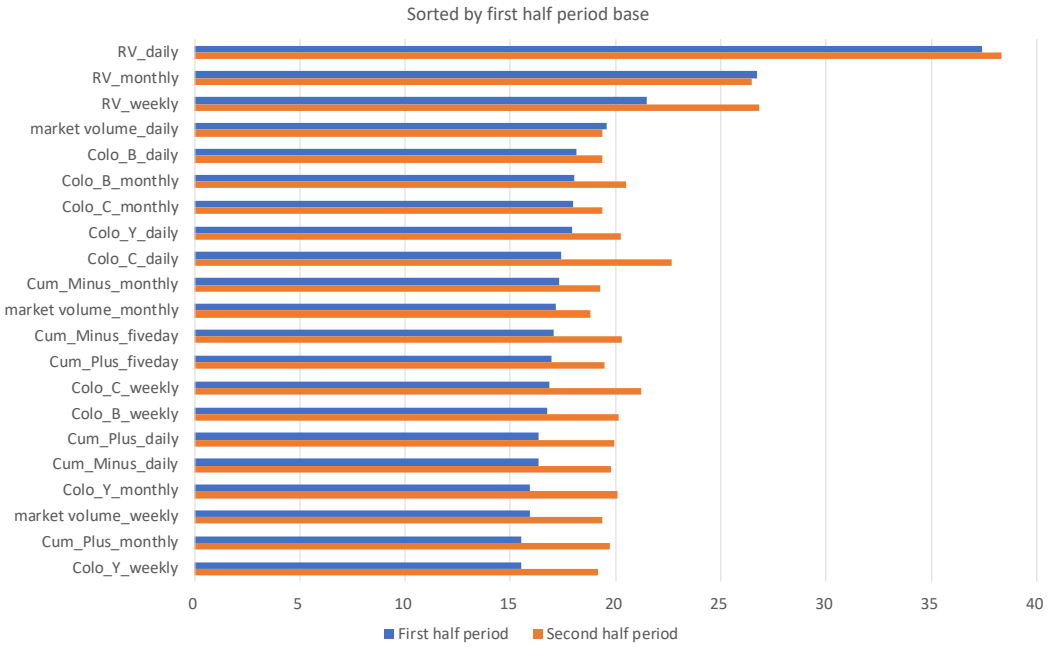

Figure 4. Importance variable changes from the first half period to the second half period sorted by first-half period base.

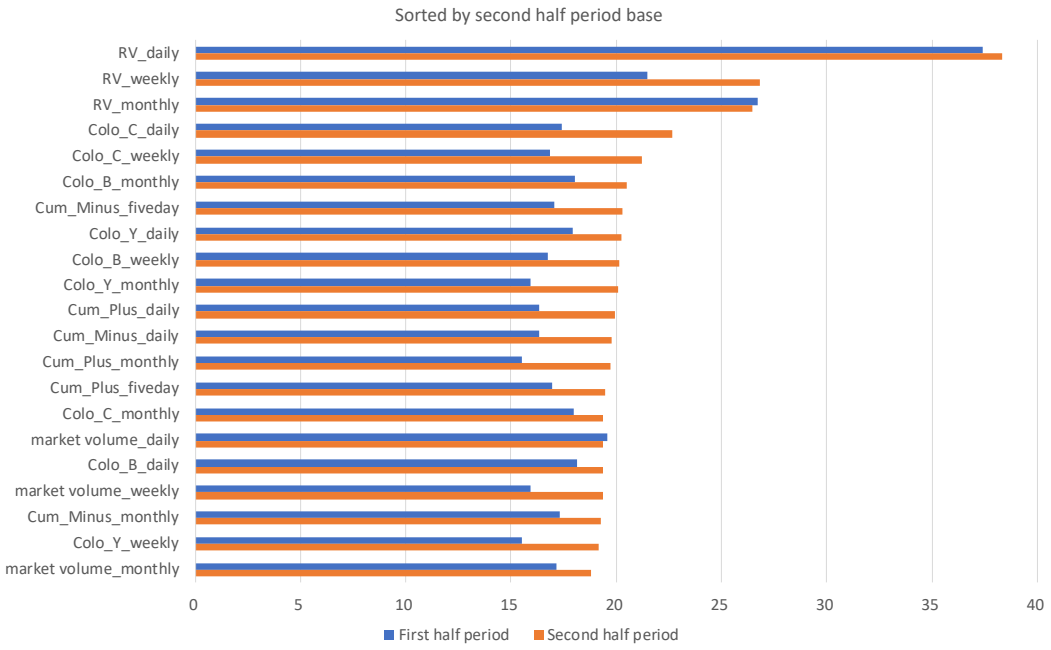

Figure 5. Importance variable changes from the first half period to the second half period sorted by second-half period base.

We consider other interpretations in this regard. Since the beginning of Arrowhead, the proportion of HFTs in the Japanese stock market has been on the rise and the transition period is thought to have continued until around 2015. For instance, all the variables in Figure 2 from 2010 to 2015 are in an uptrend. In contrast, in the second half of the period, the patterns of these variables changed. Colo_C continued to increase after 2015. This is because system renewal helps market participants to place orders more quickly. Thus, the importance of Colo_C increased during the second half period. Except for Colo_C, the others only move within this range. Although the quality of each variable may charge slightly more or less due to the effect of the Arrowhead system revision in 2015, the effect should not be as large as before and after the implementation of Arrowhead in 2010. As seen in Figure 2, the system revision in 2015 was not enough to generate a trend in Colo_B and Colo_Y. Therefore, these two TSE Co-Location variables may be less important in the second half period.

On the whole, the importance of our proposed datasets and the method of its usage, which describes the summary of the data generating process in HFT, has been increasing in proportion to the increase in the importance and participants of HFT. As evident from Table 5, the HAR + Volume + TSE Co-Location + Stock full-board model in the second half period yields a 5% higher forecast accuracy. Table 6 shows the sample size of the training and test data for each period used in this analysis.

Table 5. Prediction accuracy of RV in the first half period and second-half period.

	F-Measure	
	First Half Period	**Second Half Period**
Random Forest	0.56	0.61
Logstic	0.54	0.39

Table 6. Sample size of each period.

	First Half Period		**Second Half Period**	
	Down	**Up**	**Down**	**Up**
training data	400	387	445	457
test data	44	43	52	48

Regarding other variables, it was found that the importance of market volume has decreased and Cum_minus, which observes the changes in supply and demand, has become more important. Although this is a different result from that in the previous subsection, it goes without saying that the results will change depending on the analysis period.

5. Discussion and Conclusions

This study suggests a new approach for RV forecasts of TOPIX futures. The characteristic of our model is that it uses not only the HAR dataset but also the TSE Co-Location dataset and stock full-board dataset, both of which are related to HFT and the market volume dataset based on the random forest method. We showed that our model yields a 9% higher prediction accuracy compared to the HAR model based on the logistic method in the total observation period. In addition to this novel high accuracy, we found that the TSE Co-Location dataset, which contains HFT information, tends to play a more important role and affects daily RV forecasts. This finding is consistent with the previous studies (Zhang 2010; Haldane 2011; Benos and Sagade 2012; Caivano 2015; Myers and Gerig 2015; Kirilenko et al. 2017; Malceniece et al. 2019). In addition, this implies that transactions in high-frequency regions have an effect on economic agents who conduct transactions at a low frequency. High-frequency data are generally considered to be used only for HFT, but we have shown that it is also beneficial to the general frequency area.

These results indicate that our proposed datasets contribute to the increase in the accuracy of prediction and have a high affinity with the random forest method. The random forest method excels in prediction accuracy compared with the logistic model on average, as we expected, corresponding to the model accuracy in the previous section. Regarding this point, we consider that it is natural and consistent with previous research (Tanaka et al. 2018a; Tanaka et al. 2019), which have similar frameworks. As noted in the previous section, linear models do not work in the large explanatory variable space. In fact, from Table 3, it is evident that the logistic method seems to overfit. Practitioners and researchers should select nonlinear methods, such as the random forest method. The quality of the Japanese stock market changed due to the introduction of Arrowhead in 2010 and the effect continued strongly. Moreover, we found important rolls of the TSE Co-Location dataset deeply related to RV. We consider that the dataset associated with HFT, such as the TSE Co-Location dataset, is expected to play an important role in the model building process with the upcoming revolutions in trading systems. Now, we cannot evade high-dimensional explanatory space by adding a new dataset to the HAR. We propose the use of new datasets. By combining them with the random forest method, we show novel experimental results regarding RV forecast during two crucial events: the introduction and revision of Arrowhead. Overall, our proposed model is superior to the HAR model and can be expected to yield a high prediction accuracy. If there is a similar dataset to the TSE Co-Location dataset, our framework can be applied to the other stock markets as well.

Nevertheless, the forecasting problem may vary depending on conditions such as the selection of sampling periods. In particular, when the quality of the market changes due to new regulations, even if the model has a high accuracy of forecast in the past, in some cases, it may be necessary to revise the model in anticipation of the upcoming data in the near future. In such cases, practitioners will need to search for similar events in history, grasp the strengths and weaknesses of pattern recognition of the dataset at that time and correct them. Despite the model yielding a high prediction accuracy during the training period, it can be completely useless during the test period. Dividing the sample period appropriately, which is a universal problem, is an issue in this study. In future work, we would like to find a way to divide it automatically and conveniently.

There are pros and cons to HFT not only in Japan but also globally. Linton and Mahmoodzadeh (2018) indicate that fast algorithmic transactions place unexpectedly large orders due to program errors and algorithms that behave differently than the programmer tend to cause chain reactions, increase market volatility and disrupt market order (For instance, the May 2010 Flash Crash, August 2012 wrong order by Night Capital, October 2018 Tokyo Stock Exchange Markets Arrowhead system trouble triggered by Merrill Lynch, September 2020 Tokyo Stock Exchange Markets Arrowhead system trouble and so on). On the contrary, IOSCO reports that there is a close relationship between liquidity and volatility, in the sense that more liquidity can better absorb shocks to stock prices. HFT involved in official market-making businesses may help mitigate volatility in the short hours of the day by providing liquidity. In fact, HFT is thought to be responsible for more than 40% of the trading volume in the Japanese equity market in 2019, as shown in Figure 2. From the perspective of market liquidity and pursuit of alpha by hedge funds through HFT, we consider that HFT will play a more important role in any case. Our proposed dataset contributes to improving prediction accuracy through various types of experiments. We have experimentally shown that the degree of influence has increased in recent years. We believe that this trend will intensify in the future. Recently, the Financial Services Agency of Japan has been strengthening monitoring and legal systems, such as requiring frequent registration to trade. Along with this, the environment of HFT in the Japanese stock market is getting better, with further system improvement efforts by the Tokyo Stock Exchange, development of private exchanges and dark pools in securities firms and upgrade of securities firm systems to connect customers and securities firms more quickly.

In future work, we would like to examine this in more detail by decomposing the effect of each variable on the RV forecast improvement. Furthermore, considering Baillie et al. (2019), there may be room to extend our model, taking into account the long memory process. Chen et al. (2018) and Ma et al. (2019) are one of the helpful existing researches to expand our random forest based model to integrate long short-term memory process. In addition, we would like to extend our framework to higher-order moments. High-order moments are an important research area in finance. Hollstein and Prokopczuk (2018) showed that volatility affects stock returns. Amaya et al. (2015) insist that high-order moments are beneficial information for asset price modeling. Hence, we believe that verifying the effectiveness of our approach even in high-order moments would be helpful for both practitioners and researchers.

Author Contributions: Formal analysis, T.H.; investigation, T.H.; data curation, T.H.; writing–original draft preparation, T.H.; writing–review and editing, T.H.; supervision, K.T., T.K. and S.H.; funding acquisition, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKENHI Grant Number (A) 17H00983.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data source of our study is Japan Exchange Group and Nikkei Needs.

Acknowledgments: We are grateful to Japan Exchange Group for providing the Tokyo Stock Co-Location dataset. Also, we are grateful to the editor and three anonymous reviewers for their helpful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest. This paper gives the views of the authors, and not necessarily the position of the Nissay Asset Management.

References

Alpaydin, Ethem. 2014. *Introduction to Machine Learning*. London: The MIT Press.

Amaya, Diego, Peter Christoffersen, Kris Jacobs, and Aurelio Vasquez. 2015. Does realized skewness predict the cross-section of equity returns? *Journal f Financial Economics* 118: 135–67. [CrossRef]

Andersen, Torben G., and Tim Bollerslev. 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39: 885–905. [CrossRef]

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys. 2001. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* 96: 42–55. [CrossRef]

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys. 2003. Modeling and forecasting realized volatility. *Econometrica* 71: 529–626. [CrossRef]

Andersen, Torben G., Tim Bollerslev, and Francis X. Diebold. 2007. Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *Review of Economics and Statistics* 89: 701–20. [CrossRef]

Baillie, Richard T., Tim Bollerslev, and Hans O. Mikkelsen. 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 74: 3–30. [CrossRef]

Baillie, Richard T., Fabio Calonaci, Dooyeon Cho, and Seunghwa Rho. 2019. Long memory, realized volatility and heterogeneous autoregressive models. *Journal of Time Series Analysis* 40: 609–28. [CrossRef]

Barndorff-Nielsen, Ole E., and Neil Shephard. 2002. Estimating quadratic variation using realized variance. *Journal of Applied Econometrics* 17: 457–77. [CrossRef]

Bekaert, Geert, and Marie Hoerova. 2014. The VIX, the variance premium, and stock market volatility. *Journal of Econometrics* 183: 181–92. [CrossRef]

Benos, Evangelos, and Satchit Sagade. 2012. *High-Frequency Trading Behaviour and Its Impact on Market Quality: Evidence from the UK Equity Market*. BoE Working Paper No. 469. London: Bank of England.

Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 21: 307–28. [CrossRef]

Bollerslev, Tim, Andrew Patton, and Rogier Quaedvlieg. 2016. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics* 192: 1–18. [CrossRef]

Breiman, Leo. 2001. Random forests. *Machine Learning* 45: 5–32. [CrossRef]

Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. Classification and regression trees. In *Monterey: Wadsworth*. London: Chapman & Hall.

Caivano, Valeria. 2015. The Impact of High-Frequency Trading on Volatility. Evidence from the Italian Market. CONSOB Working Papers No. 80. Available online: https://ssrn.com/abstract=2573677 (accessed on 4 April 2021).

Chen, Zhen, Ningning He, Yu Huang, Wen Tao Qin, Xuhan Liu, and Lei Li. 2018. Integration of a deep learning classifier with a random forest approach for predicting malonylation sites. *Genomics, Proteomics & Bioinformatics* 16: 451–59.
Corsi, Fulvio. 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7: 174–96. [CrossRef]
Croft, W. Bruce, Donald Metzler, and Trevor Strohman. 2010. *Search Engines: Information Retrieval in Practice*. Boston: Addison-Wesley, p. 310.
Ding, Zhuanxin, Clive W. J. Granger, and Robert F. Engle. 1993. A Long Memory Property of Stock Market Returns and a New Model. *Journal of Empirical Finance* 1: 83–106. [CrossRef]
Engle, Robert F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987–1007. [CrossRef]
Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle. 1993. On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance* 48: 1779–802. [CrossRef]
Haldane, Andy. 2011. The race to zero. Paper presented at International Economic Association Sixteenth World Congress, Beijing, China, July 4–8.
Hansen, Peter R., and Asger Lunde. 2005. A realized variance for the whole day based on intermittent high-frequency data. *Journal of Financial Econometrics* 3: 525–54. [CrossRef]
Harvey, Andrew C. 1998. Long Memory in Stochastic Volatility. In *Forecasting Volatility in Financial Markets*. Edited by Stephen Satchell and John Knight. Amsterdam: Elsevier.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2008. *The Elements of Statistical Learning*. New York: Springer.
Hollstein, Fabian, and Mrcel Prokopczuk. 2018. How aggregate volatility-of-volatility affects stock returns. *Review of Asset Pricing Studies* 8: 253–92. [CrossRef]
Iwaisako, Tokuo. 2017. Nihon ni okeru kohindo torihiki no genjo ni tsuite (Current Status of High-Frequency Trading in Japan). Japan Securities Dealers Association. Available online: https://www.jsda.or.jp/about/iwaisakoronbun.pdf (accessed on 17 August 2020). (In Japanese).
Japan Exchange Group Connectivity Services. 2021. Available online: https://www.jpx.co.jp/english/systems/connectivity/index.html (accessed on 24 February 2021).
Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance* 72: 967–98. [CrossRef]
Linton, Oliver, and Soheil Mahmoodzadeh. 2018. Implication of high-frequency trading for security markets. *Annual Review of Economics* 10: 237–59. [CrossRef]
Luong, Chuong, and Nikolai Dokuchaev. 2018. Forecasting of realised volatility with the random forests algorithm. *Journal of Risk and Financial Management* 11: 61. [CrossRef]
Ma, Yillin, Ruizhu Han, and Xiaoling Fu. 2019. Stock prediction based on random forest and LSTM neural network. Paper presented at 19th International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea, October 15–18; pp. 126–30.
Malceniece, Laura, Kārlis Malcenieks, and Tālis J. Putniņš. 2019. High frequency trading and comovement in financial markets. *Journal of Financial Economics* 134: 381–99. [CrossRef]
Motegi, Kaiji, Xiaojing Cai, Shigeyuki Hamori, and Heifeng Xu. 2020. Moving average threshold heterogeneous autoregressive (MAT-HAR) models. *Journal of Forecasting* 39: 1035–42. [CrossRef]
Müller, Ulrich A., M. Michel Dacorogna, Rakhal D. Davé, Richard B. Olsen, Oliver V. Pictet, and Jacob E. von Weizsäcker. 1997. Volatilities of different time resolutions: Analyzing the dynamics of market components. *Journal of Empirical Finance* 4: 213–39. [CrossRef]
Myers, Benjamin, and Austin Gerig. 2015. Simulating the synchronizing behavior of high-frequency trading in multiple markets. In *Financial Econometrics and Empirical Market Microstructure*. Cham: Springer, pp. 207–13.
Nelson, Daniel B. 1991. Conditional Heteroskedasticity in Asset Returns: A new approach. *Econometrica* 59: 347–70. [CrossRef]
NIKKEI Media Marketing NEEDS Tick Data. n.d. Available online: https://www.nikkeimm.co.jp/service/detail/id=.317 (accessed on 24 February 2021).
Ohlson, James A. 1980. Financial ratios and the probabilistic prediction on bankruptcy. *Journal of Accounting Research* 18: 109–31. [CrossRef]
Patterson, Josh, and Adam Gibson. 2017. *Deep Learning: A Practitioner's Approach*, 1st ed. Newton: O'Reilly Media, Inc., p. 39.
Poon, S. Huang, and Clive W. J. Granger. 2003. Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature* 41: 478–539. [CrossRef]
Qiu, Yue, Xinyu Zhang, Tian Xie, and Shangwei Zhao. 2019. Versatile HAR model for realized volatility: A least-square model averaging perspective. *Journal of Management Science and Engineering* 4: 55–73. [CrossRef]
Shirata, Yoshiko C. 2003. Predictors of Bankruptcy after Bubble Economy in Japan: What Can You Learn from Japan Case? Paper presented at 15th Asian-Pacific Conference on International Accounting Issues, Thailand, November 1.
Tanaka, Katsuyuki, Takuji Kinkyo, and Shigeyuki Hamori. 2016. Random forests-based early warning system for bank failures. *Economics Letters* 148: 118–21. [CrossRef]
Tanaka, Katsuyuki, Takuo Higashide, Takuji Kinkyo, and Shigeyuki Hamori. 2018a. Forecasting the vulnerability of industrial economic activities: Predicting the bankruptcy of companies. *Journal of Management Information and Decision Sciences* 20: 1–24.

Tanaka, Katsuyuki, Takuji Kinkyo, and Shigeyuki Hamori. 2018b. Financial hazard map: Financial vulnerability predicted by a random forests classification model. *Sustainability* 10: 1530. [CrossRef]

Tanaka, Katsuyuki, Takuo Higashide, Takuji Kinkyo, and Shigeyuki Hamori. 2019. Analyzing industry-level vulnerability by predicting financial bankruptcy. *Economic Inquiry* 57: 2017–34. [CrossRef]

Taylor, Stephen J. 1982. Financial returns modeled by the products of two stochastic processes, a study of daily sugar prices 1961–1979. In *Time Series Analysis: Theory and Practice 1*. Edited by Oliver Duncan Anderson. Amsterdam: North-Holland, pp. 203–26.

The Japanese Government Financial Services Agency. 2018. Available online: https://www.fsa.go.jp/en./regulated/hst/index.html (accessed on 1 April 2021).

Toriumi, Fujio, Hirokazu Nishioka, Toshimitsu Umeoka, and Kenichiro Ishii. 2012. Analysis of the market difference using the stock board. *The Japanese Society for Artificial Intelligence* 27: 143–50. (In Japanese). [CrossRef]

Ubukata, Masato, and Toshiaki Watanabe. 2014. Market variance risk premiums in Japan for asset predictability. *Empirical Economics* 47: 169–98. [CrossRef]

Watanabe, Toshiaki. 2020. Heterogeneous Autoregressive Models: Survey with the Application to the Realized Volatility of Nikkei 225 Stock Index. *Hiroshima University of Economics, Keizai Kenkyu* 42: 5–18. (In Japanese).

Zhang, Frank. 2010. High-Frequency Trading, Stock Volatility, and Price Discovery. *Social Science Research Network*. Available online: http://ssrn.com/abstract=1691679 (accessed on 4 April 2021). [CrossRef]