JPX Working Paper Special Report

Expansion of the Heterogeneous Autoregressive Model with Tokyo Stock Exchange Co-Location Dataset

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; kinkyo@econ.kobe-u.ac.jp; hamori@econ.kobe-u.ac.jp

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. In this study, we propose a new framework for forecasting the realized volatility (RV) direction (“up” or “down”) of Tokyo stock price index (TOPIX) futures. Also, this study analyzes the importance of the Tokyo Stock Exchange Co-Location dataset (TSE Co-Location dataset) to forecast the RV of TOPIX futures. This paper summarizes Higashide et al. (2021), and reports the results of the analysis briefly.

1. Research Background

Andersen and Bollerslev (1998) propose using realized volatility (RV) as a proxy variable for true volatility. Watanabe (2020) remarks that the heterogeneous autoregressive (HAR) model, which is introduced by Corsi (2009), 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.

Iwaisako (2017) reports that high-frequency trading (HFT) has become an essential function in the stock markets of developed countries since the latter half of the 2000s. There are some previous studies to examine the relationship between high-frequency traders (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.

This material was compiled based on the results of research and studies by directors, officers, and/or external researchers of Japan Exchange Group, Inc., its subsidiaries, and affiliates (hereafter collectively the “JPX group”) with the intention of seeking comments from a wide range of persons from academia, research institutions, and market users. The content and opinions in this material are the authors’ own and do not constitute the official view of the JPX group or any organization to which the authors belong.

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 provides valuable information on HFT taken by the participants who trade via a server located in the TSE Co-Location area. To the best of our knowledge, this study is the first to use the TSE Co-Location dataset. However, it should be noted that not all trades via a co-location server are HFT, because co-locations are used for various purposes beyond low latency trading, such as for raising system availability in consideration of Business Continuity Plan. Thus, we utilize the dataset to construct a proxy variable that captures the effect of HFT.

2. The framework of our analysis

Firstly, we expand the HAR model using the TSE Co-Location dataset, stock full-board dataset, and market volume dataset. We build models forecasting the RV direction ("up" or "down") of Tokyo stock price index futures based on the random forest method, which is a popular machine learning algorithm and a nonlinear model. To evaluate the prediction accuracy, we use the F-measure, which is the harmonic mean that summarizes the effectiveness of precision and sensitivity in a single number, and compare the performance against conventional logistic based models

Secondly, we analyze the important variables in each model. We measure the importance of each variable by Gini index.

3. The Dataset

We use the dataset shown in Table 1. 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. The definitions of each variable are as follows:

We use the NEEDS Tick Data File provided by NIKKEI Media Marketing for RV calculation and stock full-board dataset preprocessing. Before calculation and preprocessing, we thin out every 5 min. 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.

- Realized Volatility
  Given return data $r_t, r_{t+1/n}, \dots, r_{t+(n-1)/n}$ 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

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

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.

- 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 $Bid_t, Bid_{t+1/n}, \dots, Bid_{t+(n-1)/n}$ and $Offer_t, Offer_{t+1/n}, \dots, Offer_{t+(n-1)/n}$.

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

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

We use the TSE Co-Location Dataset provided by Japan Exchange Group for TSE Co-Location and market volume.

- TSE Co-Location Dataset
  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.

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

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

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

- Market Volume

We also use the total value traded quantity as a single explanatory variable in our model, which is the denominator of Equation (7).

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

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	

Source; Higashide et al (2021) Table 1.

4. The academic contributions of this paper include:

(a) We showed that our model yields a 9% higher prediction out-of-sample accuracy compared to the HAR model based on the logistic method in the total observation period (1 March 2012 to 31 October 2019).

Table 2. 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

Source; Higashide et al (2021) Table 3

Figure 1 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. 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.

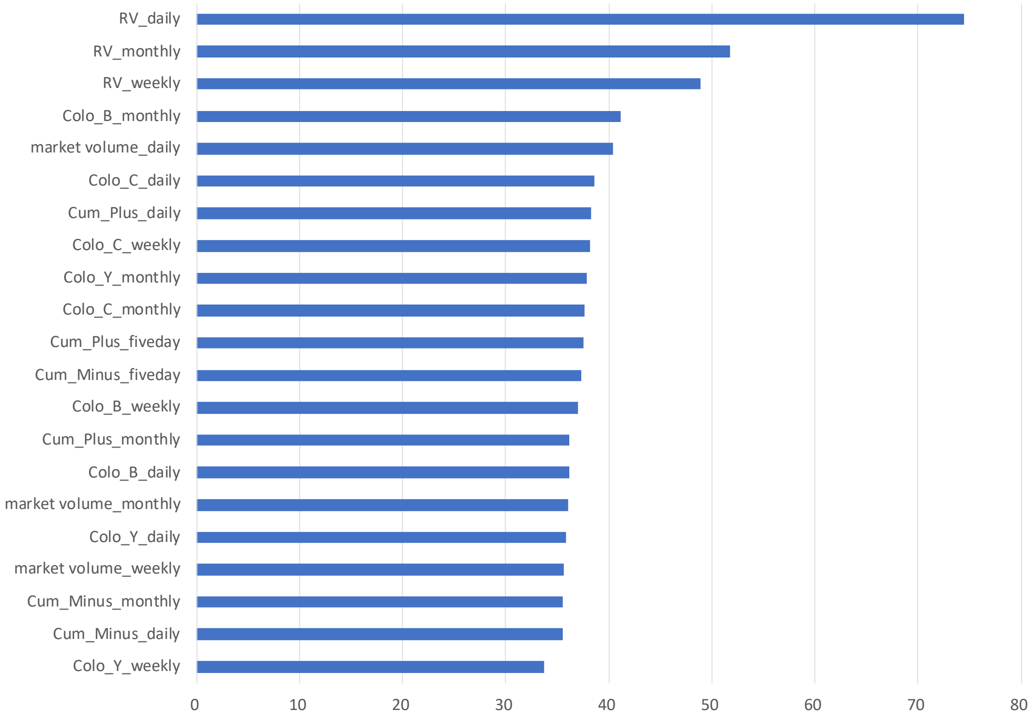

Figure 1. Important variables in the total observation period. Source; Higashide et al (2021) Figure3.

From another perspective, we look at the important variables from the time horizon: daily, weekly and monthly. Table 3 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.

Table 3. 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. Source; Higashide et al (2021) Table 4.

(b) 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 (a) to compare the differences in important variables between the first and second half periods from a time-series perspective.

Figures 2 and 3 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.

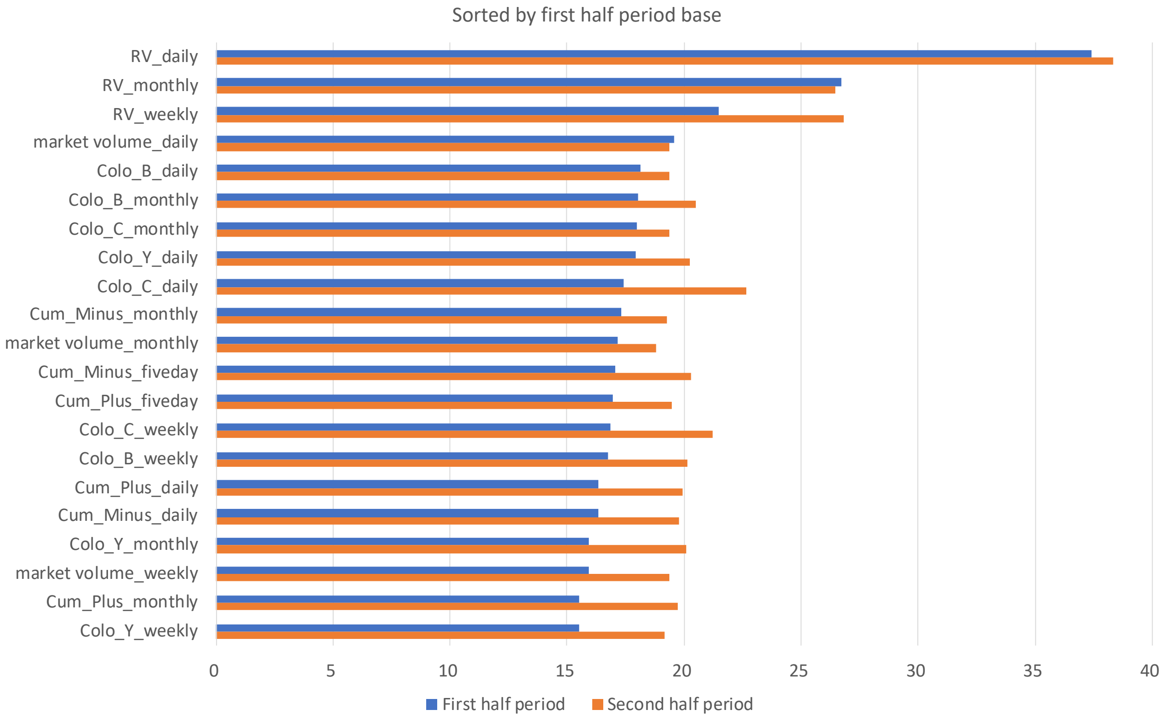

Figure 2. Importance variable changes from the first half period to the second half period sorted by first-half period base. Source; Higashide et al (2021) Figure4.

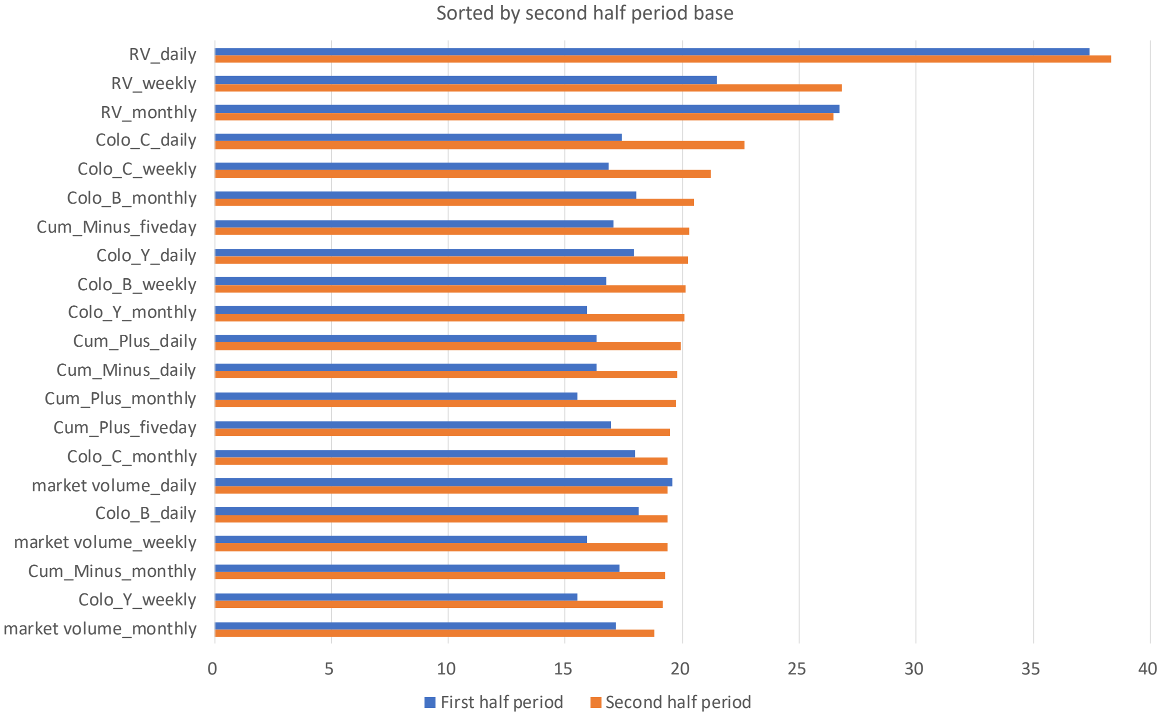

Figure 3. Importance variable changes from the first half period to the second half period sorted by second-half period base. Source; Higashide et al (2021) Figure5.

(c) We found that the random forest method framework works effectively and can be superior to the linear model in the framework of RV forecast.
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 (Table 4). These results are consistent with those of previous studies, which reported that linear models do not work where there are many explanatory variables in space.

Table 4. 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

Source; Higashide et al (2021) Table 5

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, there may be room to extend our model, taking into account the long memory process. In addition, we would like to extend our framework to higher-order moments. It is not clear whether the HFTs have an impact on volatility, or conversely, whether each HFT itself anticipates volatility and trades accordingly. Though a lot of discussion are made about pros and cons to HFT not only in Japan but also globally[1], there is no unique definition of the HFTs and the HFT. In order to make international comparisons, we think that we should consider global common standards for the definitions of the HFT and the HFTs in the future.

Reference

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

(Benos and Sagade 2012) 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

(Caivano 2015) 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).

(Corsi 2009) Corsi, Fulvio. 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7: 174–96.

(Haldane 2011) Haldane, Andy. 2011. The race to zero. Paper presented at International Economic Association Sixteenth World Congress, Beijing, China, July 4–8.

(Hansen and Lunde 2005) 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.

(Higashide et al. 2021) Higashide T, Tanaka K, Kinkyo T, Hamori S. 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. 2021; 14: 215.

(Iwaisako 2017) 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)

(Kirilenko et al. 2017) 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.

(Malceniece et al. 2019) Malceniece, Laura, Kārlis Malcenieks, and Tālis J. Putniņš. 2019. High frequency trading and comovement in financial markets. *Journal of Financial Economics* 134: 381–99.

(Myers and Gerig 2015) 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.

(Watanabe 2020) 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 2010) 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).

[1] Some system troubles are caused by program error of HFT, such as wrong order by Night Capital on August 2012. Of course, there are troubles happened regardless from HFTs, such as Tokyo Stock Exchange Markets trouble on October 2020.