University of Pretoria
Department of Economics Working Paper Series

Forecasting Realized Stock-Market Volatility: Do Industry Returns have Predictive Value?

Riza Demirer
Southern Illinois University Edwardsville
Rangan Gupta
University of Pretoria
Christian Pierdzioch
Helmut Schmidt University

Working Paper: 2020-107
December 2020

Department of Economics
University of Pretoria
0002, Pretoria
South Africa
Tel: +27 12 420 2413

Forecasting realized stock-market volatility: Do industry returns have predictive value?

Riza Demirer[a], Rangan Gupta[b], Christian Pierdzioch[c]

December 2020

Abstract

Yes, they do. Utilizing a machine-learning technique known as random forests to compute forecasts of realized (good and bad) stock market volatility, we show that incorporating the information in lagged industry returns can help improve out-of sample forecasts of aggregate stock market volatility. While the predictive contribution of industry level returns is not constant over time, industrials and materials play a dominant predictive role during the aftermath of the 2008 global financial crisis, highlighting the informational value of real economic activity on stock market volatility dynamics. Finally, we show that incorporating lagged industry returns in aggregate level volatility forecasts benefits forecasters who are particularly concerned about under-predicting market volatility, yielding greater economic benefits for forecasters as the degree of risk aversion increases.

JEL classification: G17; Q02; Q47

Keywords: Stock market; Realized volatility; Industry returns, Market efficiency and information

[a] Corresponding author. Department of Economics & Finance, Southern Illinois University Edwardsville, Edwardsville, IL 62026-1102; Email: rdemire@siue.edu.

[b] Department of Economics, University of Pretoria, Pretoria, 0002, South Africa; Email address: rangan.gupta@up.ac.za.

[c] Department of Economics, Helmut Schmidt University, Holstenhofweg 85, P.O.B. 700822, 22008 Hamburg, Germany; Email address: c.pierdzioch@hsu-hh.de.

Acknowledgements: CP thanks the Deutsche Forschungsgemeinschaft for financial support (Project: Exploring the experience-expectation nexus in macroeconomic forecasting using computational text analysis and machine learning; Project number: 275693836).

1 Introduction

Volatility forecasting is a key component of option pricing, hedging and portfolio optimization applications. Naturally, there exists a large strand of literature that offers a wide-array of univariate and multivariate models to forecast stock market volatility (e.g. Granger and Poon, 2003; Engle and Rangel, 2008; Rapach et al., 2008; Rangel et al., 2011; Engle et al., 2013, Ben Nasr et al., 2016, and Salisu et al. 2020 for a detailed discussion of research in this area). Despite the multitude of studies on stock market volatility forecasting using a wide range of predictors that include macroeconomic and financial variables, however, the literature has not yet examined the predictive power of industry level information over aggregate level stock market volatility. This paper adds to this line of research by investigating for the first time the role of lagged industry returns from across the entire economy in predicting aggregate stock market volatility. Indeed, we show that incorporating the information in lagged industry returns can help improve out-of-sample forecasts of aggregate stock market volatility, rendering significant economic benefits for a forecaster, particularly as the degree of risk aversion increases.

In a well cited study, Hong et al. (2007) present the theoretical framework towards the predictive power of industry returns for stock market return forecasting. According to the so-called gradual diffusion of information hypothesis, the information contained in industry returns diffuses gradually across markets as a result of the interaction of boundedly rational investors with access to private information at different points in time. In this setting, public information gets partially reflected in asset prices such that certain types of investors, such as those who specialize in trading the broad market index, experience a lag in receiving industry level information that is already accessible to investors who specialize in particular industries. This, in turn, forms the basis for return predictability at the aggregate market level as industry level dynamics contain predictive information regarding the economic fundamentals that lead the aggregate stock market. Although later studies including Hong et al. (2014), Tse (2015), Ciner (2019) and Rapach et al. (2019) present conflicting evidence regarding the predictive content of industry returns, interestingly, the literature has not yet extended the analysis to stock market volatility forecasting. To the best of our knowledge, ours is the first to examine the predictive power of lagged industry returns over aggregate stock market volatility.

In our forecasting application, we use a machine-learning technique known as random forests (Breiman 2001) to compute forecasts of realized (good and bad) stock-market volatility. Random forests have been used in recent applications to study the predictive value of industry returns for stock market returns (Ciner, 2019) and the realized volatility of intraday Bitcoin returns (Bouri et al., 2020). In our case, instead of relying on conditional volatility models from the generalized autoregressive conditional heteroskedasticity (GARCH)-family, we follow Andersen and Bollerslev (1998) and forecast monthly realized volatility (RV) as measured by the sum of squared daily log-returns over a month. The use of realized volatility provides an observable measure of the latent process of volatility that is model-free unlike the conditional estimates of the same.

As for the econometric framework is concerned, we utilize the popularly employed heterogeneous autoregressive realized-volatility model (HAR-RV) of Corsi (2009) that allows to capture stylized facts such as multi-scaling behavior and long-memory of the volatility process in a straightforward and simple way. Although the ordinary-least-squares technique is commonly applied to estimate the HAR-RV model, the use of random forests in our forecasting application has several advantages. First, random forests render it possible to analyze the links between realized volatility and a large number of predictors (in our case, lagged industry returns from across the entire economy) in a fully data-driven way. Second, random forests automatically capture potential nonlinear links between realized volatility and its predictors as well as any interaction effects among the predictors. Finally, unlike the ordinary-least-squares technique, random forests always yield forecasts of realized volatility that are non-negative.

While the empirical findings confirm the superior forecasting performance of random forests over the classic HAR-RV model, we also find that lagged industry returns indeed contain valuable predictive information over the aggregate stock market realized volatility as well as its “good” (upward) and “bad” (downward) variants, both under the standard symmetric and an asymmetric loss functions. Incorporating lagged industry returns in volatility forecasting models benefits particularly a forecaster who suffers more from an under-prediction than an over-prediction, while these benefits tend to decrease with the length of the forecast horizon. We further show that out-of-sample realized volatility forecasts that incorporate the information in lagged industry returns are economically valuable, with greater economic benefits for a forecaster at higher levels of risk aversion. Finally, we show that certain industries including those that reflect real economic activity play a more dominant role across the short and long horizons than others, which is in line with the gradual diffusion of information as opposed to an efficient market setting.

The remainder of the paper is structured as follows. In Section 2, we describe how a random forest is grown and present our data. In Section 3, we report the empirical results, followed by a discussion of the economic implications of our findings in Section 4. Finally, in Section 5, we conclude the paper with final remarks.

2 Methodology and Data

2.1 Random forests

A random forest is an ensemble machine-learning technique, consisting of a large number of individual regression trees (for a textbook exposition, see Hastie et al. 2009; our notation follows theirs). A regression tree, T, consists of branches that subdivide the space of predictors, $\mathbf{x} = (x_1, x_2, ...)$, of realized volatility (in our case) into l non-overlapping regions, R_l. These regions are formed by applying a search-and-split algorithm in a recursive top-down fashion.

Starting at the top level of a regression tree, the algorithm iterates over the predictors, s, and the corresponding splitting points, p, that can be formed using the data on a predictor. For every combination of a predictor and a splitting point, the algorithm computes two half-planes, $R_1(s,p) = \{x_s | x_s \leq p\}$ and $R_2(s,p) = \{x_s | x_s > p\}$. The search for an optimal combination of a predictor and a splitting point minimizes the standard squared-error loss criterion:

$$\min_{s,p} \left\{ \min_{\bar{RV}_1} \sum_{x_s \in R_1(s,p)} (RV_i - \bar{RV}_1)^2 + \min_{\bar{RV}_2} \sum_{x_s \in R_2(s,p)} (RV_i - \bar{RV}_2)^2 \right\}, \tag{1}$$

where the index i identifies those data on realized volatility that belong to a half-plane, and $\bar{RV}_k = \text{mean}\{RV_i \, | x_s \in R_k(s,p)\}, k = 1,2$ denotes the half-plane-specific mean of realized volatility. The outer minimization searches over all combinations of s and p. Given s and p, the inner minimization minimizes the half-plane-specific squared error loss by an optimal choice of the half-plane-specific means of realized volatility. The solution of the minimization problem given in Equation (1) yields the top-level optimal splitting predictor, the top-level optimal splitting point, and the two region-specific means of realized volatility. Accordingly, the solution yields a first simple regression tree that has two terminal nodes.

At the next stage, the minimization problem in Equation (1) is solved separately for the two optimal top-level half-planes, $R_1(s,p)$ and $R_2(s,p)$, in order to grow a larger regression tree. The new solution yields up to two second-level optimal splitting predictors and optimal splitting points, and four second-level region-specific means of realized volatility. Upon repeating this search-and-split algorithm multiple times, we are able to grow an increasingly complex regression tree. Finally, the search-and-split algorithm stops when a regression tree has a preset maximum number of terminal nodes or every terminal node has a minimum number of observations. We use a cross-validation approach to identify the optimal minimum number of observations per terminal node (see Section 3.1 for further details).

When the search-partition algorithm stops, the regression tree sends the predictors from its top level to the various leaves along the various optimal partitioning points (nodes) and branches. We then use the regression tree to forecast realized volatility by its region-specific mean. For a regression tree made up of L regions, we compute forecasts as follows (**1** denotes the indicator function):

$$T\left(\mathbf{x}_i, \{R_l\}_1^L\right) = \sum_{l=1}^{L} \bar{RV}_l \mathbf{1}(\mathbf{x}_i \in R_l). \tag{2}$$

While the search-and-split algorithm can be used in principle to compute finer and finer granular forecasts of realized volatility, the resulting growing complexity of the hierarchical structure of a regression tree gives rise to an overfitting and data-sensitivity problem, which, in turn, deteriorates the forecasting performance. A random forest solves this problem as follows. First, a large number of bootstrap samples (sampling with replacement) is obtained from the data. Second, to each bootstrap sample, a random regression tree is fitted. A random regression tree differs from a standard regression tree in that the former uses for every splitting step only a random subset of the predictors, which mitigates the effect of influential predictors on tree building. Growing a large number of random trees decorrelates the forecasts from individual trees, and averaging the decorrelated forecasts obtained from the individual random regression trees stabilizes the forecasts of realized volatility.

2.2 Data

We use monthly excess returns for 49 value-weighted industry portfolios for the period January 1946- December 2019, obtained from Ken French's online data library.[1] Following the convention, we exclude"others" and end up with 48 industries defined based on the Standard Industrial Classification (SIC) system. Separately, daily and monthly stock market returns are collected as the returns of a value-weighted market portfolio from the Center for Research in Security Prices (CRSP). Daily market returns are used to compute the realized market volatility estimates (RV) for each month from log daily returns (r_t) as follows:

$$RV_t = \sum_{i=1}^{N} r_i^2, \tag{3}$$

where N denotes the number of data available for the month. In addition to realized volatility, we examine "good" and "bad" realized volatility. The categorization of RV into its good and bad components is an important issue as Giot et al. (2010) stresses that financial market participants care not only about the level of volatility, but also of its nature, with all traders making the distinction between good and bad volatilities. The "good" and "bad" components of realized volatility are formulated as the upside and downside realized semi-variances (RV^B and RV^G), respectively, computed from positive and negative returns (see Barndorff-

[1] Available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

Nielsen et al., 2010) as follows:

$$RV_t^B = \sum_{i=1}^{T} r_i^2 \, I_{[(r_i)<0]}, \tag{4}$$

$$RV_t^G = \sum_{i=1}^{T} r_i^2 \, I_{[(r_i)>0]}. \tag{5}$$

The model to forecast realized volatility follows the widely-employed heterogeneous autoregressive realized volatility (HAR-RV) model of Corsi (2009) that has now become one of the most popular models in the empirical finance literature on modeling and forecasting realized volatility. The theoretical foundation of the HAR-RV model is laid out by the so-called heterogeneous market hypothesis of Müller et al. (1997). The heterogeneous market hypothesis stipulates that the stock market is populated by different types of traders who differ with respect to their sensitivity to information flows at different time horizons. In this setting, market participants with short- versus long-term investment horizons respond to information flows heterogeneously at different time horizons.

Accordingly, the key idea underlying the HAR-RV model is to use realized volatilities from different time resolutions as predictors of realized volatility. When studying daily realized volatility, it is common practice among researchers to consider daily, weekly, and monthly realized volatilities as predictors of subsequent realized volatility. In our case, because we study monthly data, in line with the strand of the literature that deals with the lead-lag relationship between industry and aggregate level returns, we model the month-h-ahead realized volatility, RV_{t+h}, using the current realized volatility, RV_t, the quarterly realized volatility, $RV_{t,q}$, computed as the average realized volatility from month $t-3$ to month $t-1$, and the yearly realized volatility, $RV_{t,y}$, computed as the average realized volatility from month $t-12$ to month $t-1$. We compute these quarterly and yearly average realized volatilities for the standard measure of realized volatility and for good and bad realized volatility.

– Please include Figure 1 about here. –

Figure 1 presents the time series plots of computed realized volatility, RV_t, series along with its low-frequency components, $RV_{t,q}$, and $RV_{t,y}$, in Panel A, and the corresponding counterparts for bad and good realized volatility in Panels B and C. We observe notable spikes in the realized volatility estimates around the stock market crash of 1987 and later during the 2008 global financial crisis period. Comparing Panels B and C, we observe that bad realized volatility was the dominant factor in the case of the 1987 stock market crash, while the 2008 global financial crisis period was equally plagued by both the good and bad components of realized volatility.

3 Empirical analysis

3.1 Calibration

We use rolling-estimation windows of length 120, 180, 240, and 360 months to estimate both the baseline HAR-RV model (the model that excludes lagged industry returns) and the HAR-RV model extended to include lagged industry returns. We forecast realized volatility one-month, three-months and one-year ahead (that is, we set $h = 1, 3, 12$) by estimating random forests in the statistical computing program R (R Core Team 2019) using the add-on package "grf" (Tibshirani et al. 2020). While shifting the rolling-estimation windows across the data set, we optimize, by means of cross validation, the number of predictors randomly selected for splitting, the minimum node size of a tree, and the parameter that governs the maximum imbalance of a node. We optimize these parameters separately for the baseline HAR-RV model and the extended HAR-RV model that features lagged industry returns.[2] We use 2,000 random regression trees to grow a random forest.

3.2 The classic HAR-RV model as a benchmark

In order to set the stage for our empirical analysis, it is useful to go back for the moment to the classic HAR-RV model. In the context of our analysis, the classic HAR-RV model is formulated as $RV_{t+h} = \beta_0 + \beta_1 RV_t + \beta_2 RV_{t,q} + \beta_3 RV_{t,y} + \varepsilon_t$, where $\beta_j,\ j = 0,1,2,3$ are the coefficients to be estimated by means of the ordinary-least-squares (OLS) technique, ε_t is an error term, and RV_{t+h} denotes the realization of realized volatility in month $t+h$. The classic HAR-RV model extended to include lagged industry returns is then formulated as $RV_{t+h} = \beta_0 + \beta_1 RV_t + \beta_2 RV_{t,q} + \beta_3 RV_{t,y} + \sum_{j=1}^{48} \beta_{j+3} r_{t,j} + \varepsilon_t$. The corresponding random-forest models can, thus, be expressed as $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$ when we exclude lagged industry returns, and as $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$ when we include lagged industry returns in the array of predictors. At this point, it is worth noting that our framework ensures that (i) random forests do not necessarily invoke a linear structure as does the OLS technique, and (ii) random forests go beyond the OLS technique in that they allow the predictors (lagged industry returns in our case) to interact in an arbitrary data-driven way.

We compare in Table 1, for various rolling-estimation windows and forecast horizons, the out-of-sample performance of the HAR-RV model estimated by means of the OLS technique with the out-of-sample performance of random forests in terms of the root-mean-squared error (RMSE) statistics implied by these two models. To this end, we estimate both models by excluding lagged industry returns (that is, the set of predictors includes $RV_t, RV_{t,q}, RV_{t,y}$ only; the OLS model also

[2] The "grf" package also allows different subsamples to be used for constructing a tree and for making predictions. We deactivate this option, as in a classic random forest.

features a constant) and then by including lagged industry returns. We then compute the RMSE statistics for both models and compute the corresponding ratios. A ratio larger than unity indicates that the random forests outperform the corresponding HAR-RV model in terms of the RMSE statistic. Finally, we repeat these calculations for good and bad volatility.

– Please include Table 1 about here. –

Three main results emerge from Table 1. First, the RMSE ratio exceeds unity in the vast majority of cases, indicating the superior forecasting performance of random forests against the OLS. Second, when we exclude lagged industry returns from the set of predictors, we observe a RMSE ratio smaller than unity only for the short forecast horizon and when we consider a relatively long rolling-estimation window of 360 observations, while the random forests outperform the OLS in most other cases even when lagged industry returns are not included in the model. Third and more importantly, the RMSE ratios are found to be substantially larger when we include lagged industry returns in the set of predictors. In other words, the results show that industry level information can indeed improve the out-of-sample performance of forecasting models for aggregate stock market realized volatility, while random forests systematically outperform the standard HAR-RV model estimated by the OLS technique.

The observed superior performance of the random forest against the OLS is not unexpected given that the HAR-RV model extended to include lagged industry returns from across the entire economy requires the estimation of many parameters. In case some of these industries have only limited predictive power for realized volatility, their estimated parameters will add noise to the forecasts of realized volatility. This brings about a trade-off when the OLS technique is used to estimate the HAR-RV model as the improvement in the forecast performance due to the predictive power of industry returns has to be weighed against a deteriorated forecast performance due to an overparameterization of the model. Random forests, in contrast, do not suffer from such an overparameterization problem as the search-and-split algorithm that is used to grow regression trees automatically discards less informative predictors in a data-driven way as we recursively subdivide the predictor space into rectangular non-overlapping arrays in a top-down fashion.

Finally, we follow a slightly more sophisticated modeling approach by adding only one of the 48 lagged industry returns at a time to the classic HAR-RV model, estimating the resulting 48 models by the OLS technique and finally averaging the forecasts from these 48 models to forecast realized volatility. When we compare the forecasts computed by means of random forests with the forecasts obtained from such a thick-modeling approach, once again, we find that random forests yield superior performance in terms of the RMSE statistic (results are not reported but are available from the authors upon request).

3.3 The predictive power of lagged industry returns

Having established evidence of superior forecasting performance of random forests agains the OLS, we summarize in Table 2, for the four rolling-estimation windows and the three forecast horizons studied, the ratios of the RMSEs of the restricted, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$, and the full random-forest model that includes lagged industry returns, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$. Panels B and C present the results for good and bad volatility. The ratios in Table 2 exceed unity for the majority of cases, indicating that the full model that incorporates lagged industry returns outperforms the restricted model for realized volatility as well as its bad and good variants. This means that incorporating industry level information in the forecasting model helps to improve the accuracy of out-of-sample forecasts of aggregate stock market volatility. The magnitude of the ratios of the RMSEs tends to be larger for the short forecast horizon than for the two longer forecast horizons, especially for realized volatility and good realized volatility, suggesting that industry level information can be particularly useful to improve relatively shorter term stock market volatility forecasts and for bullish market states.

– Please include Table 2 about here. –

Because large forecast errors have a disproportionately large effect on the RMSE statistic, we report in Table 3 the results obtained from the ratio of the mean-absolute errors (MAE) statistic of the restricted model and the full model. Again, a value larger than unity for this ratio indicates that lagged industry returns improve forecast accuracy. The results in Table 3 corroborate those for the RMSE ratios reported in Table 2. We find that the full model that incorporates lagged industry returns outperforms the restricted model in the majority of cases with generally larger MAE ratios observed at the short forecast horizon, further supporting the predictive value of industry level information particularly for shorter horizon volatility forecasts and for bullish market states.

– Please include Table 3 about here. –

In volatility forecasting exercises that involve noisy volatility proxies, Patton (2010) shows that the quasi-likelihood (QLIKE) loss function along with the usual mean-squared-error loss function allow for an unbiased model ordering. Therefore, in order to check the robustness of our findings, we report in Table 4 the results for the popular QLIKE loss function. We observe that the QLIKE ratios are smaller than for the RMSE and MAE statistics, however still larger than unity except for some cases for $h = 12$. Moreover, the QLIKE ratios tend to become larger when the length of the rolling-estimation window increases. These additional results, thus, further confirm the predictive value of lagged industry returns over stock market volatility forecasts irrespective of the loss function utilized to assess forecast accuracy.

In a recent application of loss functions to volatility forecasting, Reschenhofer et al. (2020) propose two alternative likelihood-based loss functions, one based on a t-distribution (QLIKE-t) and the other based on an F-distribution (QLIKE-F), that are less sensitive to outliers and thus allow for a more stable ranking of forecasts. Given this evidence, we also experimented with the QLIKE-t and QLIKE-F distributions and found qualitatively similar results (for alternative degrees-of-freedom parameters) to those obtained from QLIKE loss function.[3]

– Please include Table 4 about here. –

Finally, having confirmed the predictive value of lagged industry returns via alternative loss functions, as another approach, we report in Table 5 the results of Clark and West (2007) test of equality of mean-squared prediction errors of the full model that includes lagged industry returns and the restricted model that excludes industry level information. The test yields significant results at the 5% and, in a few cases, at the 10% level of significance at the two shorter forecast horizons for realized volatility and its good and bad variants, confirming that the full model outperforms the restricted model in most cases. While the test statistic takes on smaller values for the long forecast horizon, it remains significant in the majority

[3] In order to save space, we do not report the results for the QLIKE-t and QLIKE-F distributions but make them available upon request.

of cases at the 10% level of significance, and in few cases even at the 5% percent level of significance.[4] Overall, various methods to assess the predictive value of lagged industry returns yield consistent findings confirming that incorporating industry level information in forecasting models can improve the out-of-sample accuracy of stock market volatility forecasts.

– Please include Table 5 about here. –

3.4 Time-varying importance of industry returns

In their popularly cited study, Hong et al. (2007) show that 14 out of 34 industries, including commercial real estate, petroleum, metal, retail, financial, and services, can predict stock market movements by one month, while other industries including petroleum, metal, and financials can forecast the market even two months ahead. However, re-examining these results with updated data, Tse (2015) shows that only one to seven industries have significant predictive ability for the stock market. Although these studies focus on stock market return forecasting and rely on in-sample tests, they bring about an interesting question as to the time-varying importance of industry returns in aggregate level market forecasts. For this reason, we supplement our analysis by examining the relative importance of the predictors over time in order to see whether certain industries play a more prominent predictive role in our volatility forecasts.

We present in Figure 2 the relative importance of the predictors in the full model, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$, measured in terms of how often a

[4] We also examine, by means of the Clark-West test, how random forests perform relative to a HAR-RV model estimated by means of the OLS technique, where both models feature lagged industry returns as predictors in addition to the standard HAR-RV predictors. The results (not reported but available upon request) corroborate those reported in Table 1 and show that the forecasts computed by means of random forests significantly outperform the OLS forecasts.

predictor is used for splitting when building a tree. Given the large number of industries used in the set of predictors, in order to ease the interpretation of the results, we aggregate the data into an “rv” block that represents the components of the HAR-RV model, and eleven broad industry groups (energy, materials, industrials, consumer staples, consumer discretionary, healthcare, financials, IT, communication, utilities, real estate).

– Please include Figure 2 about here. –

As expected, given the popularity of the HAR-RV model in empirical finance, the three terms of the HAR-RV model (treated in the figure as a single block) always play an important role in the volatility forecasts. Interestingly, however, the predictive role of the “rv” block that represents the components of the HAR-RV model, changes over time. We observe that the predictive role of the “rv” block has gained momentum during the period preceding the Global Financial Crisis (GFC) of 2008 and then peaked during the GFC, suggesting that the importance of non-industrial information including behavioral factors and/or changes in investors’ risk aversion increased during the run up to the global crash. However, we also observe at the short and intermediate forecast horizons, an increasing role of industrials and materials during the aftermath of the global crisis, highlighting the informational value of real economic activity on stock market realized volatility forecasts. Interestingly, at the long forecast horizon, we observe a similar pattern for consumer related industries with consumer discretionary and consumer staples taking on a greater role in the predictive models. Overall, our analysis suggests that certain industries play a more dominant predictive role in aggregate level volatility forecasts and the predictive contribution of industry level returns is not constant over time with a structural change occurring during the period that precedes the global financial crisis.

3.5 Robustness checks

In order to further confirm the inferences discussed so far, we report in this section the findings from a battery of robustness checks. In Table 6, we summarize the results of the Clark-West test when we add market returns as a control variable to the array of predictors of the full model. Specifically, we test the null hypothesis that a model that features the standard HAR-RV terms, lagged market returns, and lagged industry returns has the same out-of-sample forecasting performance as a model that features only the standard HAR-RV terms and lagged market returns. The results are qualitatively similar to those reported in Table 5, suggesting that industry returns indeed capture significant predictive information for subsequent realized market volatility over and above lagged market returns.

– Please include Table 6 about here. –

Table 7 reports the results of four additional robustness checks (for the sake of brevity, we focus on realized volatility). First, we replace the rolling-estimation window by means of a recursively expanding estimation window. Second, we forecast, for $h > 1$, the average realized volatility formulated as $\text{mean}(RV_{t+1} + \ldots + RV_{t+h})$. Third, we forecast the realized standard deviation. These additional robustness checks lend further support to our conclusion that industry returns capture valuable predictive information for subsequent realized market volatility.

– Please include Table 7 about here. –

As a fourth and final robustness check, we consider boosted regression trees (Friedman 2001, 2002) as an alternative to random forests. Boosted regression trees combine regression trees with elements of statistical boosting. They resemble random forests insofar as the key idea is to grow a forest of trees by combining simple regression trees. In contrast to random forests, however, boosted regression trees are estimated by means of a forward stage-wise iterative algorithm. The specific algorithm that we consider is known as stochastic gradient-descent boosting. Estimating the stochastic gradient-descent variant of boosted regression trees using the R add-on package “gbm” (Greenwell et al. 2020) and applying the Clark-West test to the estimated forecasts, we observe significant results further confirming the predictive value of lagged industry returns, particularly for shorter forecast horizons.

3.6 Asymmetric loss and quantile-random forests

The results reported in the preceding sections are based on the assumption that a forecaster’s loss is a symmetric function of the squared or absolute forecast error (with the QLIKE loss function being the exception). That is, an under-prediction of realized volatility causes the same loss as an over-prediction of the same size. In practical settings, however, one could easily think of situations in which the loss function of a forecaster, such as one who uses certain options-trading strategies, is asymmetric in the forecast error. Therefore, in order to account for such a setting, we assume that a forecaster has the following loss function (e.g., Elliott et al. 2005):

$$L(FE_{t+h},\alpha) = [\alpha + (1-2\alpha)I_{[FE_{t+h}<0]}]|FE_{t+h}|^{p}. \tag{6}$$

where we compute the forecast error, FE, by subtracting the forecast of realized volatility from the actual realization of realized volatility. Special cases of this loss function are the lin-lin loss function ($p = 1$) and the quad-quad loss function ($p = 2$) where the shape parameter $\alpha \in (0,1)$ determines the asymmetry of the loss function. The loss function is symmetric for the special case $\alpha = 0.5$. Hence, the parameter configuration $\alpha = 0.5$ and $p = 1$ implies that a forecaster’s loss is a symmetric function of the mean-absolute forecast error, while $\alpha = 0.5$ and $p = 2$ implies that a forecaster's loss is a symmetric function of the squared forecast error. Setting the shape parameter to $\alpha > 0.5$, in turn, results in a loss function that attaches a higher loss to an under-prediciton of realized volatility than to an over-prediction of the same (absolute) size. In the opposite case, $\alpha < 0.5$, an over-prediction is costlier than a corresponding under-prediction.

We assume that a forecaster whose loss function is asymmetric forecasts a quantile of the conditional distribution of realized volatility that corresponds to the shape of the loss function rather than simply the mean (or median) of realized volatility. Specifically, we assume that a forecaster who has a loss function with a shape parameter $\alpha > 0.5$ (that is, a forecaster who suffers a higher loss from an under-prediction than from an over-prediction) adjusts his or her forecast upward. Such a forecaster, thus, forecasts a quantile of the conditional distribution of realized volatility above the median. Conversely, a forecaster who has a loss function with a shape parameter $\alpha < 0.5$ adjusts his or her forecast downward relative to the median. We compute such upward- and downward-adjusted forecasts by estimating quantile-random forests (Meinshausen 2006).

We proceed as follows. We estimate quantile-random forests to compute forecasts of the $\alpha-$quantiles of the conditional distribution function of realized volatility and then compute the forecast errors that correspond to the estimated $\alpha-$quantiles. We use the resulting forecast errors to compute a forecaster's loss according to the loss function given in Equation (6). Cumulating the losses over all out-of-sample forecasts for both the restricted model, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$, and the full model, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$, we finally compute the ratio of the cumulated losses, where a loss ratio that exceeds unity indicates that the full model produces a lower cumulated loss than the restricted model.

– Please include Figure 3 about here. –

Figure 3 presents the plots for the loss ratios for $p = 1$ and $p = 2$ as a function of the shape parameter, α. We average for each shape parameter, α, the loss ratios across the four different forecast horizons. The loss ratios obtained in this way are found to be larger than unity for all three forecast horizons when a forecaster's loss is a function of the lin-lin type ($p = 1$), except for shape parameters close to its upper boundary. For a loss function of the quad-quad type ($p = 2$), the loss ratio is smaller than unity for some shape parameters smaller than $\alpha = 0.5$, and for shape parameters close to its upper boundary. In contrast, for a broad range of shape parameters above 0.5, the loss ratio is found to be larger than unity with the loss ratio attaining a maximum in this range for both the lin-lin and the quad-quad types. This indicates that a forecaster who suffers more from an under-prediction than an over-prediction (and whose loss function has a shape parameter not too close to its upper boundary) reaps relatively larger benefits from incorporating lagged industry returns to stock market volatility forecasting models as well as models of good and bad realized volatility. The benefits, however, tend to decrease with the length of the forecast horizon, and are particularly large in the case of good realized volatility and $h = 1$. Nevertheless, these additional tests provide additional insight to the potential benefits a forecaster can achieve by incorporating industry level information to generate volatility forecasts at the aggregate market level.

4 Economic implications

Considering alternative shapes of a forecaster's loss function is one way to quantify the economic benefits of forecasts, and the discussion in the Section 3.6 indicates that incorporating lagged industry returns in aggregate level volatility forecasts benefits forecasters who are particularly concerned about under-predicting market volatility. This is certainly an important consideration for the pricing of options contracts as ignoring industry level information can potentially lead to under-pricing of these securities. An alternative way to assess the economic implications of our findings is to directly use a forecaster's utility function to measure the benefits from utilizing industry level information in forecasting realized market volatility. To this end, we consider a forecaster who uses forecasts to decide whether to invest in a portfolio consisting of a riskless asset and in the stock market. Like Cenesizoglu and Timmermann (2012), we keep things simple in that we abstract from transaction costs, neglect intertemporal hedging considerations and focus on the short forecast horizon ($h = 1$). We assume that a forecaster has a constant relative risk aversion utility function (CRRA; that is $U(W) = W^{1-\gamma}/(1-\gamma)$, where U denotes utility and W denotes wealth), for which we consider three levels of risk aversion ($\gamma = 3, 5, 10$) similar to the application by Cenesizoglu and Timmermann (2012). We further fix the riskless interest rate at zero and assume that a forecaster uses the returns as observed in the period in which a forecast is to be formed to forecast returns. We then compute the certainty equivalent returns (CER) for a forecaster who utilizes lagged industry returns to forecast realized market volatility and, alternatively, for a forecaster who ignores industry level information.

– Please include Table 8 about here. –

In Table 8, we report the difference (in percent) between the resulting CER values for the two types of forecasters. A positive number in the table indicates that a forecaster attains a higher CER by incorporating lagged industry returns in the forecasting model. The results for realized volatility in Panel A indicate significant economic gains from using industry level information when forecasting realized market volatility (only one figure in the table is negative). The same also holds for good realized volatility in Panel C with the magnitude of the economic benefits from utilizing lagged industry returns increasing as the degree of risk aversion of a forecaster increases. This suggests that more risk averse forecasters can reap increasingly greater economic benefits from incorporating industry level information in aggregate level volatility forecasts. In the case of bad realized volatility, in contrast, the results are mixed. A longer rolling-estimation window tends to worsen the economic value added by industry returns to the forecasting model. This could be due to the dominance of non-industry related factors such as behavioral and sentiment related effects over stock market volatility dynamics, particularly during periods of market crisis when investors would be more likely to engage in herding behavior. Nevertheless, our results indicate that a forecaster who plans to use lagged industry returns to forecast bad realized market volatility should choose a relatively short rolling-estimation window, especially in case he or she is highly risk averse.

5 Concluding Remarks

In a well-cited study, Hong et al. (2007) argue that industry portfolios capture predictive information over the aggregate stock market, in line with the so-called gradual diffusion of information hypothesis that suggests the information contained in industry returns diffuses gradually across markets. Although later studies provide mixed evidence regarding the predictive power of industry returns over stock market returns, the literature has not yet examined the predictability of stock market volatility in this context. Given the importance of accurate volatility forecasts for a number of financial activities including option pricing, hedging and portfolio optimization, this paper adds to this line of research by investigating for the first time the role of lagged industry returns from across the entire economy in predicting aggregate stock market volatility.

Utilizing a machine-learning technique known as random forests to compute forecasts of realized (good and bad) stock-market volatility, we show that incorporating the information in lagged industry returns can indeed help improve out-of-sample forecasts of aggregate stock market volatility. The predictive contribution of industry level returns, however, is not constant over time with an increasing role of industrials and materials during the aftermath of the 2008 global financial crisis, highlighting the informational value of real economic activity on stock market volatility dynamics. We also show that incorporating incorporating lagged industry returns in aggregate level volatility forecasts benefits forecasters who are particularly concerned about under-predicting market volatility. Finally, assuming a constant relative risk aversion utility function, we show that the magnitude of the economic benefits from utilizing lagged industry returns increases as the degree of risk aversion of a forecaster increases, suggesting that more risk averse forecasters can reap increasingly greater economic benefits from incorporating industry level information in their stock market volatility forecasts.

As a final note, it is important to point out that the purpose of our empirical analysis is not to show that random forests and related tree-based techniques are the best machine-learning techniques for forecasting realized volatility. Many different techniques populate the machine-learning zoo and the comparative advantages of these techniques can be used to shed light on various different aspects of realized volatility. The purpose of our empirical analysis is to shed light on the potential role of industry returns for forecasting stock market volatility at the aggregate level and random forests turn out to be particularly useful in this regard. For the purpose of our analysis, random forests have several advantages. They are straightforward to implement and produce results that are easy to interpret. They can easily manage a large number of industries from across the entire economy as predictors of realized volatility and they always produce nonnegative forecasts of realized volatility. In future research, it will be interesting to build on our empirical results by applying other machine-learning techniques to examine whether industries lead the stock market in the context of return and volatility prediction.

References

Andersen T.G., and Bollerslev T. (1998). Answering the skeptics: yes, standard volatility models do provide accurate forecasts. International Economic Review, 39(4): 885−905.

Barndorff-Nielsen, O.E., Kinnebrouk, S., and Shephard, N. (2010). Measuring downside risk: realised semivariance. Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle, (Edited by T. Bollerslev, J. Russell and M. Watson), Oxford University Press: 117−136.

Ben Nasr, A. Lux, T., Ajmi, A.N., and Gupta, R. (2016). Forecasting the volatility of the Dow Jones Islamic stock market index: Long memory vs. regime switching. International Review of Economics and Finance, 45(1), 559−571.

Breiman, L. (2001). Random forests. *Machine Learning*, 45: 5−32.

Bouri, E., Gkillas, K., Gupta, R., and Pierdzioch, C. (2020). Forecasting realized volatility of Bitcoin: The role of the trade war. *Computational Economics*, forthcoming.

Cenesizoglu, T., and Timmermann, S. (2012). Do return prediction models add economic value?. *Journal of Banking and Finance*, 36: 2974−2987.

Ciner, C. (2019). Do industry returns predict the stock market? A reprise using the random forest. *Quarterly Review of Econmics and Finance*, 72: 152−158.

Clark, T.D., and West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138: 291−311.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7: 174−196.

Elliott, G., Komunjer, I., and Timmermann, A. (2005). Estimation and testing of forecasting rationality under flexible loss. *Review of Economic Studies*, 72: 1107−1125.

Engle, R.F., and Rangel, J.G. (2008). The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Causes. Review of Financial Studies 21(3): 1187−1222.

Engle, R.F., Ghysels, E., and Sohn, B. (2013). Stock Market Volatility and Macroeconomic Fundamentals. The Review of Economics and Statistics 95(3): 776−797.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29: 1189−1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38: 367−378.

Giot, P., Laurent, S., and Petitjean, M. (2010). Trading activity, realized volatility and jumps.Journal of Empirical Finance, 17(1): 168−175.

Greenwell, B., Boehmke, B., Cunningham, J. and GBM Developers (2020). gbm: Generalized boosted regression models. R package version 2.1.8. https://CRAN.R-project.org/package=gbm.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York, NY: Springer.

Hong, H., Lim, T. and Stein, J.C. (2000). Bad news travels slowly: size, analyst coverage and the profitability of momentum strategies. *Journal of Finance*, 55: 265−295.

Hong, H., Torous, W., Valkanov, R. (2007). Do industries lead stock markets? *Journal of Financial Economics*, 83: 367−396.

Hong, H., Torous, W., Valkanov, R., (2014). Note on “Do industries lead stock markets?”. `http://rady.ucsd.edu/docs/faculty/valkanov/Note_10282014.pdf`.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning*, 7: 983−999.

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., and Pictet, O. V. (1997). Volatilities of different time resolutions − Analyzing the dynamics of market components. *Journal of Empirical Finance*, 4: 213−239.

Patton, A.J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160: 246−256.

Poon, S-H, and Granger, C. W. J. (2003). Forecasting Volatility in Financial Markets: A Review. Journal of Economic Literature, 41(2): 478−539.

Rangel, J.G., and Engle, R.F. (2011). The Factor-Spline-GARCH Model for High and Low Frequency Correlations. Journal of Business & Economic Statistics 30(1): 109−124.

Rapach, D.E., Strauss, J.K., and Wohar, M.E. (2008). Forecasting stock return volatility in the presence of structural breaks, in Forecasting in the Presence of Structural Breaks and Model Uncertainty, in David E. Rapach and Mark E. Wohar (Eds.), Vol. 3 of Frontiers of Economics and Globalization, Bingley, United Kingdom: Emerald: 381−416.

Rapach, D.E. and Zhou, G. (2013). Forecasting stock returns. In: Elliott, G., and Timmermann, A. (Eds.), *Handbook of Economic Forecasting*. Volume 2A, Amsterdam: Elsevier: 328−383.

Rapach, D.E., Strauss, J.K., Tu, J., and Zhou, G. (2019). Industry Return Predictability: A Machine Learning Approach, *Journal of Financial Data Science* 1(3): 9−28.

Reschenhofer, E., Mangat, M. K., Stark, T. (2020). Volatility forecasts, proxies and loss functions. *Journal of Empirical Finance*, 59: 133−153.

Tse, Y. (2015). Do industries lead stock markets? A reexamination. *Journal of Empirical Finance*, 34: 195−203.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `URLhttps://www.R-project.org/`.

Salisu, A.A., Gupta, R., and Ogbonna, A.E. (2020). A Moving Average Heterogeneous Autoregressive Model for Forecasting the Realized Volatility of the US Stock Market: Evidence from Over a Century of Data. International Journal of Finance & Economics. DOI: `https://doi.org/10.1002/ijfe.2158`.

Tibshirani,J., Athey, S., and Wager, S. (2020). grf: Generalized Random Forests. R package version 1.1.0. `https://CRAN.R-project.org/package=grf`.

Table 1: Comparing OLS and random forests by means of root-mean-squared-error ratios

Panel A: Realized volatility

	Excluding industry returns			Including industry returns		
Window	$h=1$	$h=3$	$h=12$	$h=1$	$h=3$	$h=12$
120	1.0155	1.0514	1.0794	1.4342	1.5083	1.5878
180	1.0136	1.0361	1.0134	1.2485	1.2714	1.3162
240	0.9088	1.0378	1.0091	1.1873	1.2286	1.2687
360	0.8869	0.9926	0.9756	1.0013	1.1370	1.0840

Panel B: Bad realized volatility

	Excluding industry returns			Including industry returns		
Window	$h=1$	$h=3$	$h=12$	$h=1$	$h=3$	$h=12$
120	1.0374	1.1304	1.2616	1.4203	1.6603	1.8148
180	0.9950	1.0334	1.0689	1.2125	1.3638	1.3303
240	0.8645	1.0002	0.9849	1.1191	1.2479	1.2667
360	0.9266	0.9766	0.9926	1.0567	1.1626	1.0861

Panel C: Good realized volatility

	Excluding industry returns			Including industry returns		
Window	$h=1$	$h=3$	$h=12$	$h=1$	$h=3$	$h=12$
120	0.9559	1.0306	1.0487	1.3428	1.5160	1.6616
180	0.9481	1.0551	1.0342	1.2419	1.2766	1.3547
240	0.9780	1.0528	1.0210	1.1807	1.2279	1.2642
360	0.9371	1.0155	1.0098	1.0258	1.1337	1.0967

Note: This table reports RMSE ratios, computed for out-of-sample forecasts. The columns entitled "Excluding industry returns" compare the standard HAR-RV model, $RV_{t+h} = \beta_0 + \beta_1 RV_t + \beta_2 RV_{t,q} + \beta_3 RV_{t,y} + \varepsilon_t$, with the corresponding random-forest model, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$. The columns entitled "Including industry returns" compare the extended HAR-RV model, $RV_{t+h} = \beta_0 + \beta_1 RV_t + \beta_2 RV_{t,q} + \beta_3 RV_{t,y} + \sum_{j=1}^{48} \beta_{j+3} r_{t,j} + \varepsilon_t$, with the corresponding random-forest model, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$. A ratio larger than unity indicates that the random-forest model outperforms the corresponding HAR-RV model estimated by means of the OLS technique in terms of the RMSE criterion. The column entitled "Window" shows the length of the rolling estimation window. The parameter h denotes the forecast horizon (in months). The random forests are built using 2,000 trees.

Table 2: The predictive power of lagged industry returns (root-mean-squared-error ratios)

Panel A: Realized volatility

Window	$h=1$	$h=3$	$h=12$
120	1.1175	1.0195	0.9831
180	1.0905	1.0027	1.0427
240	1.1587	1.0050	1.0603
360	1.0553	1.0240	1.0477

Panel B: Bad realized volatility

Window	$h=1$	$h=3$	$h=12$
120	1.0183	1.0142	0.9898
180	1.0319	1.0086	1.0217
240	1.0273	0.9905	1.0481
360	1.0120	1.0266	1.0056

Panel C: Good realized volatility

Window	$h=1$	$h=3$	$h=12$
120	1.1417	1.0298	0.9942
180	1.1443	1.0030	1.0193
240	1.1214	1.0126	1.0431
360	1.1023	1.0269	1.0178

Note: This table reports the ratio of the RMSE statistics of the restricted random-forest model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$) and the full random-forest model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$) that includes lagged industry returns. The column entitled "Window" shows the length of the rolling estimation window. The parameter h denotes the forecast horizon (in months). The random forests are built using 2,000 trees.

Table 3: The predictive power of lagged industry returns (mean-absolute-error ratios)

Panel A: Realized volatility

Window	$h=1$	$h=3$	$h=12$
120	1.1232	1.0146	0.9209
180	1.1359	1.0172	0.9967
240	1.2024	1.0272	1.0435
360	1.1207	1.0616	1.0282

Panel B: Bad realized volatility

Window	$h=1$	$h=3$	$h=12$
120	1.0456	1.0052	0.9564
180	1.0294	0.9917	1.0146
240	1.0538	0.9766	1.0321
360	1.1093	1.0445	1.0094

Panel C: Good realized volatility

Window	$h=1$	$h=3$	$h=12$
120	1.1575	1.0325	0.9385
180	1.1914	1.0454	0.9887
240	1.2311	1.0842	1.0312
360	1.2047	1.0792	1.0205

Note: This table reports results the ratio of the MAE statistics of the restricted random-forest model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$) and the full random-forest model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$) that includes lagged industry returns. The column entitled "Window" shows the length of the rolling estimation window. The parameter h denotes the forecast horizon (in months). The random forests are built using 2,000 trees.

Table 4: The predictive value of lagged industry returns (QLIKE ratios)

Panel A: Realized volatility

Window	$h=1$	$h=3$	$h=12$
120	1.0023	1.0056	0.9762
180	1.0045	1.0066	0.9860
240	1.0052	1.0088	0.9985
360	1.0100	1.0155	0.9991

Panel B: Bad realized volatility

Window	$h=1$	$h=3$	$h=12$
120	0.9927	1.0097	0.9716
180	1.0034	1.0096	0.9818
240	1.0063	1.0021	0.9921
360	1.0102	1.0178	1.0147

Panel C: Good realized volatility

Window	$h=1$	$h=3$	$h=12$
120	1.0186	1.0000	0.9893
180	1.0192	1.0040	0.9958
240	1.0230	1.0082	1.0161
360	1.0293	1.0208	1.0197

Note: This table reports results the ratio of the quasi-likelihood (QLIKE) losses of the restricted random-forest model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$) and the full random-forest model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$) that includes lagged industry returns. The column entitled "Window" shows the length of the rolling estimation window. The parameter h denotes the forecast horizon (in months). The random forests are built using 2,000 trees.

Table 5: The predictive power of industry returns (Clark-West test)

Panel A: Realized volatility

Window	$h = 1$	$h = 3$	$h = 12$
120	1.5683	2.9308	0.6815
180	1.4486	3.7398	1.6944
240	1.5940	2.8703	1.3060
360	1.4922	3.2668	1.0630

Panel B: Bad realized volatility

Window	$h = 1$	$h = 3$	$h = 12$
120	2.1304	2.9925	0.7488
180	1.8272	3.1523	1.2456
240	1.5408	2.3829	1.3182
360	2.3595	1.9863	1.3193

Panel C: Good realized volatility

Window	$h = 1$	$h = 3$	$h = 12$
120	1.9868	1.9544	1.2794
180	1.7752	2.1362	1.3833
240	1.8707	2.5040	1.9851
360	2.2922	2.9637	1.9287

Note: This table reports the results of the Clark and West (2007) test of equal mean-squared prediction errors. The null hypothesis is that the full model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$) that includes lagged industry returns has the same out-of-sample forecasting performance as the restricted model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$). The alternative hypothesis is that the full model performs better than the restricted model. Results are based on Newey-West robust standard errors. Critical values are 1.28 and 1.65 at the 10% and 5% level of significance. The column entitled "Window" shows the length of the rolling estimation window. The parameter h denotes the forecast horizon (in months). The random forests are built using 2,000 trees.

Table 6: Out-of-sample tests after controlling for market returns

Panel A: Realized volatility

Window	$h = 1$	$h = 3$	$h = 12$
120	1.6663	2.9304	0.1223
180	1.2691	3.4922	1.2396
240	1.5509	3.2289	1.1011
360	0.1339	2.8273	1.0864

Panel B: Bad realized volatility

Window	$h = 1$	$h = 3$	$h = 12$
120	2.1542	2.6248	1.1641
180	2.8119	2.5963	1.5343
240	2.5513	2.0545	1.1304
360	-0.1072	1.9242	0.6938

Panel C: Good realized volatility

Window	$h = 1$	$h = 3$	$h = 12$
120	1.6526	2.1705	1.1455
180	2.0736	2.3758	1.7618
240	1.7681	2.2312	1.7680
360	0.0734	2.6230	1.7470

Note: This table reports the results of the Clark-West test. The null hypothesis is that the full model (the model that features the standard HAR-RV terms, market returns, and lagged industry returns) has the same out-of-sample forecasting performance as the restricted model (the model that features only the standard HAR-RV terms and market returns). The alternative hypothesis is that the full model performs better than the restricted model. Results are based on Newey-West robust standard errors. Critical values are 1.28 and 1.65 at the 10% and 5% level of significance. The column entitled “Window” shows the length of the rolling estimation window. The parameter *h* denotes the forecast horizon (in months). The random forests are built using 2,000 trees.

Table 7: Further robustness checks

Specification	$h=1$	$h=3$	$h=12$
Recursive estimation of random forests	1.9651	3.2487	1.5887
Forecasting average realized volatility	1.6346	1.6989	1.7393
Forecasting realized standard deviation	2.2200	4.0051	1.1863
Forecasting by boosted regression trees	1.8892	2.4335	0.8385

Note: This table reports results of the Clark-West test. The null hypothesis is that the full model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$) that includes lagged industry returns has the same out-of-sample forecasting performance as the restricted model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$). The alternative hypothesis is that the full model performs better than the restricted model. Results are based on Newey-West robust standard errors. Critical values are 1.28 and 1.65 at the 10% and 5% level of significance. The column entitled "Specification" shows the model variant being studied. Recursive estimation of random forests: A recursive estimation window replaces the rolling-estimation window. The initial training period is 120 months. Forecasting average RV: for $h > 1$ forecasts of mean$(RV_{t+1} + ... + RV_{t+h})$ are computed. For $h = 1$, the test result slightly differs from the corresponding test result reported in Table 5, reflecting random variation due to bootstrapping and random tree building. The length of the rolling estimation window is 120 months. Forecasting realized standard deviation: Forecasts are computed for the square root of RV. The random forests are built using 2,000 trees. The parameter h denotes the forecast horizon (in months). The parameters used for estimation of boosted regression trees are as follows: tree depth = 5, learning rate equal = 0.005, minimum number of observations per node 5 a bag fraction = 0.5. The maximum number of trees is 2,000. The number of trees used for estimation is determined by five-fold cross-validation.

Table 8: Economic implications

Panel A: Realized volatility

Window	$\gamma = 3$	$\gamma = 5$	$\gamma = 10$
120	2.92	8.12	36.98
180	1.87	6.29	34.08
240	-1.06	3.75	30.76
360	6.04	15.00	80.61

Panel B: Bad realized volatility

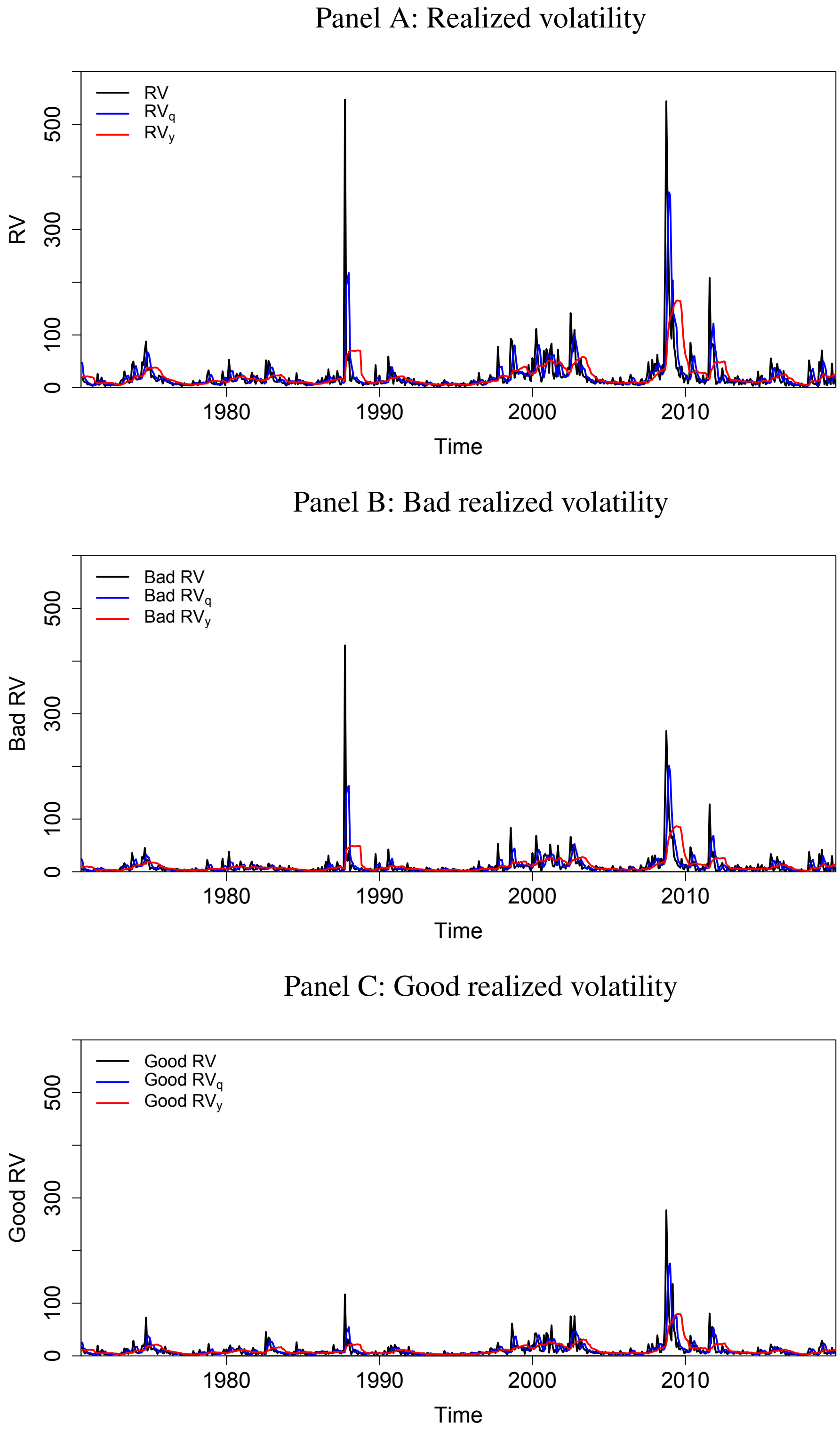
Window	$\gamma = 3$	$\gamma = 5$	$\gamma = 10$
120	0.16	1.26	10.22
180	-2.12	0.89	5.58
240	-3.50	-2.49	-7.67
360	-4.61	-6.95	-9.76

Panel C: Good realized volatility

Window	$\gamma = 3$	$\gamma = 5$	$\gamma = 10$
120	3.90	10.39	46.23
180	4.06	10.01	48.83
240	5.58	10.94	56.74
360	11.63	29.44	281.25

Note: This table reports the difference (in percent) between the certainty equivalent returns (CER) that a forecaster attains upon using lagged industry returns to forecast realized market volatility and otherwise. A positive number indicates that a forecaster attains higher CER ratio when using lagged industry returns to set up a forecasting model (that is, we compute (CER with lagged industry returns − CER without industry returns) / | CER without industry returns |). The utility function is of the constant-relative-risk-aversion (CRRA) type, where the parameter γ captures the degree of risk aversion. The risk-free interest rate is fixed at zero and there are no transaction costs. A forecaster rebalances his or her portfolio every month, where portfolio weights are restricted to the interval $[0, 1.5]$. Out-of-sample forecasts are for the full model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$) that includes lagged industry returns and the restricted model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$). The forecasting horizon is fixed at $h = 1$. The random forests are built using 2,000 trees. The maximum number of trees is 2,000. The number of trees used for estimation is determined by five-fold cross-validation.

Figure 1: The components of the HAR-RV model

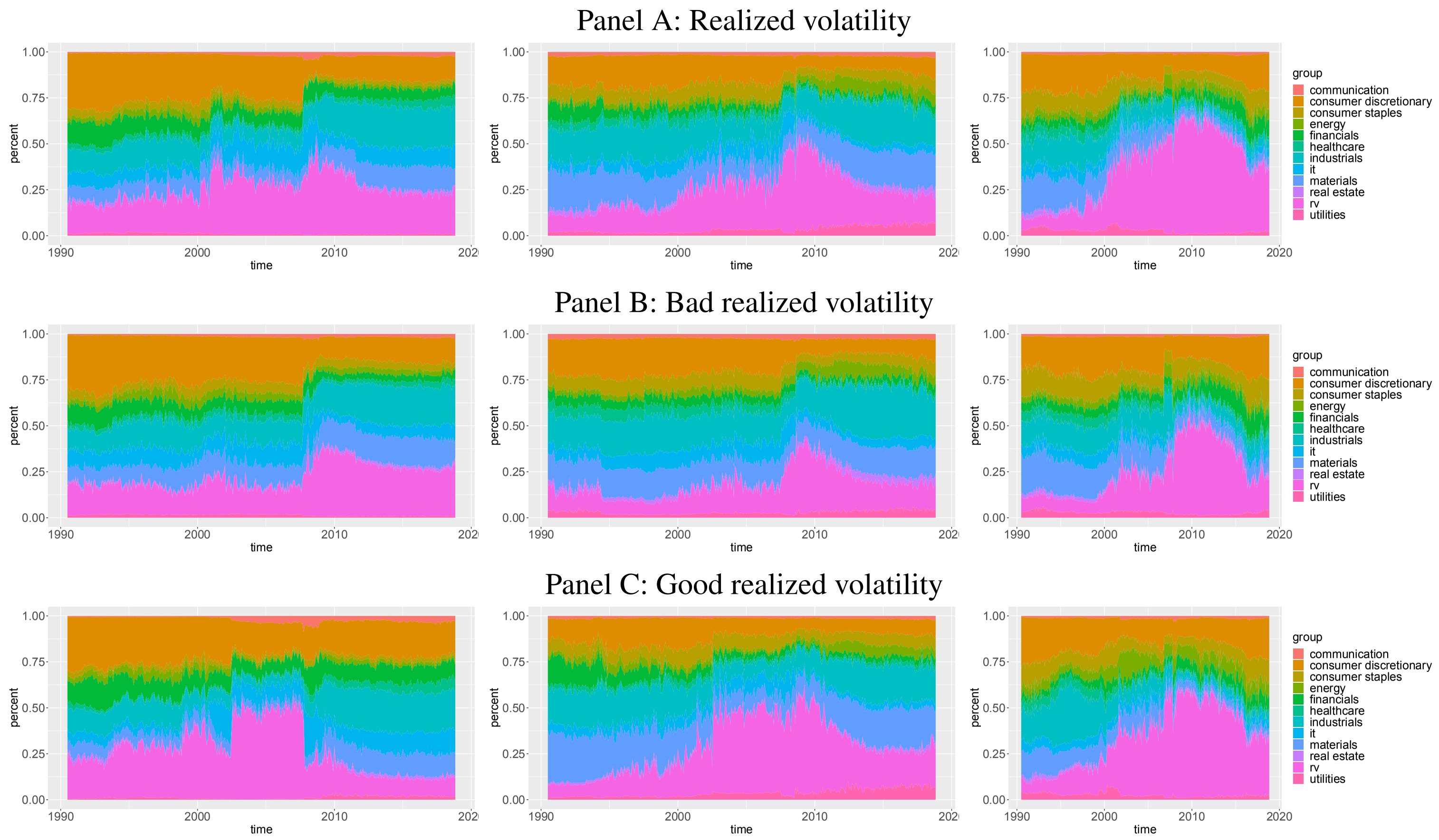


Panel A: Realized volatility

Panel B: Bad realized volatility

Panel C: Good realized volatility

Figure 2: The relative importance of predictors



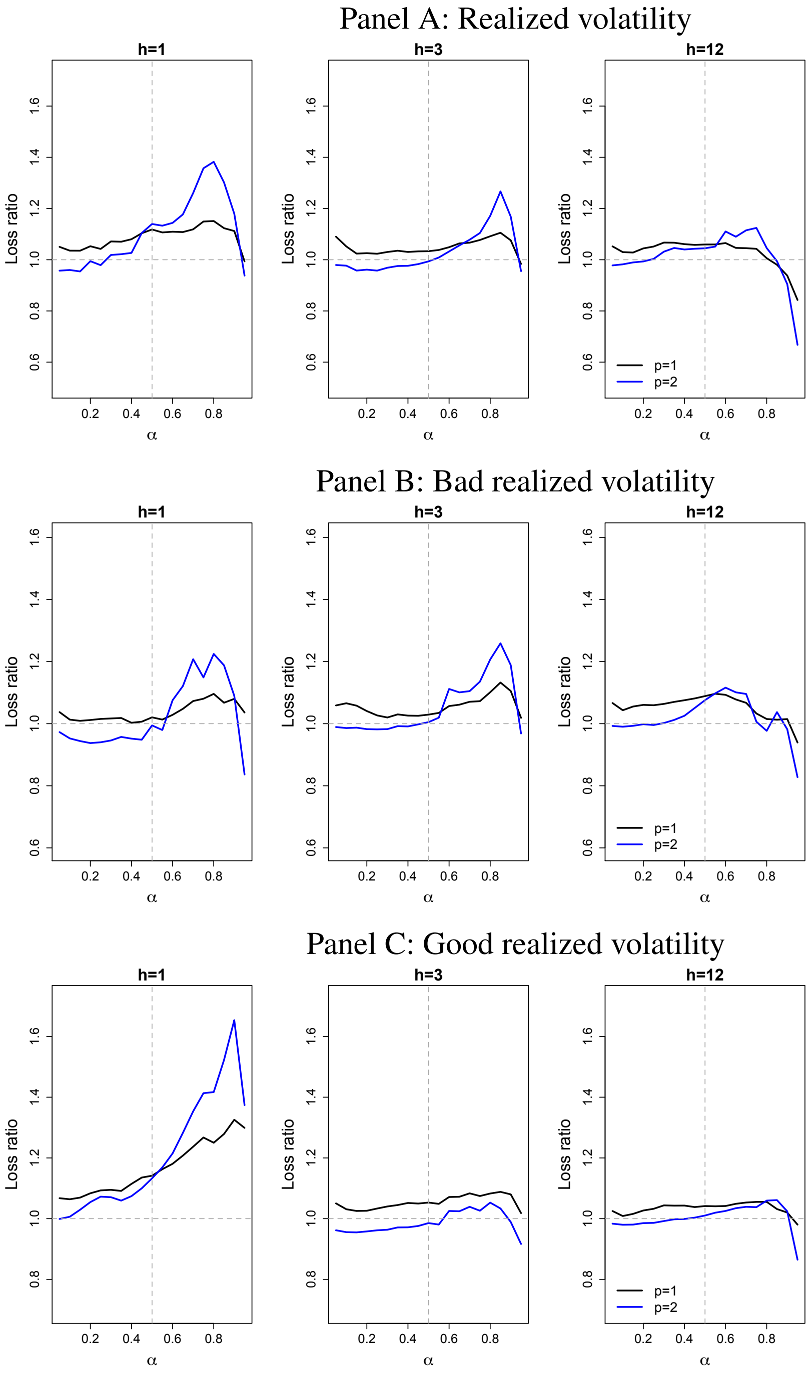
Panel A: Realized volatility

Panel B: Bad realized volatility

Panel C: Good realized volatility

Note: Predictor importance is computed for the full model ($RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, \ldots, r_{t,48})$) and a rolling-estimation window of length 240 months. Predictor importance is defined as the weighted sum of how often a predictor is used for splitting. Maximum tree depth considered: 4. Numbers are averaged across 10 estimations of the random forests. The forecasts horizons are $h = 1, 3, 12$ (from left to right). The random forests are built using 2,000 trees.

Figure 3: The shape of a forecaster's loss function and the predictive value of lagged industry returns



Panel A: Realized volatility

Panel B: Bad realized volatility

Panel C: Good realized volatility

Note: This figure displays the ratio of the cumulated loss for the restricted random-forest model, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y})$, and the full random-forest model, $RV_{t+h} = RF(RV_t, RV_{t,q}, RV_{t,y}, r_{t,1}, r_{t,2}, ..., r_{t,48})$ that includes lagged industry returns. The loss function is given in Equation 6. A ratio larger than unity signals that the full model performs better than the restricted model. The loss ratios are averaged across the four different rolling-estimation windows (120,180, 140, and 360 months). The random forests are estimated by setting the minimum node size to 5 and using one-third of the predictors randomly chosen for splitting. The random forests are built using 2,000 trees. The parameter h denotes the forecast horizon (in months).