Erasmus University

Erasmus School of Economics

Bachelor Thesis Quantitative Finance

Combining Volatility Forecasts: A Shapley Value Based Approach[1]

Hugo Groenewegen

542628hg

This paper investigates the contribution of combining volatility models based on a decomposition of the R^2 via Shapley values. Basing weights using Shapley values can be beneficial as it is prone to highly correlated independent variables. Volatility models are used in two environments, first using Monte Carlo simulations where 100 generated series are used to calculate performance metrics and conduct tests. Second, using data from 2000 up to and including 2020 on the London Stock Exchange (FTSE100). Models like GARCH and HAR are used in the analysis. The Shapley weights and their performances are compared to the arithmetic average of the forecasts and to OLS coefficients. The research shows that a Shapley value-based approach significantly outperforms the arithmetic average in standard deviations, loss functions, and Diebold Mariano tests both in the simulation and the FTSE100 data set.

Supervisor: Prof. Dr. P.H.B.F Franses

Second Assessor: Dr. W. Wang

2 July 2023

[1] The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam

Contents

1 Introduction

This past century there have been many tumultuous events, like the Great Recession in 2008 or the Covid-19 pandemic. These events did not only have a big impact on people's day-to-day life, but also the stock markets felt these events constituted big losses which eventually were recovered over time. Events like these contribute to a higher level of volatility, which is defined as the degree of variation of a trading price series over time. The question is how to forecast this volatility as well as possible. Volatility forecasting is relevant for investors as it is directly linked to calculating the risk of investments. Several forecasting models have been used in the past decades, ranging from simple to more sophisticated models. Predominately the GARCH(1,1) model [Bollerslev, 1986] has been studied intensively for the past decades as it would outperform other models in its forecasting ability [Andersen et al., 2005]. However, would it not be better to combine several models into a more weighted forecast? Furthermore, what weights would be the best? As recently stated by Franses (2023) it is beneficial to base these weights on the Shapley values of the explanatory variables based on the R^2, instead of simply taking the arithmetic average, where every forecast gets an equal weight of $\frac{1}{n}$ (n being the number of forecasts). This leads to the research question of this thesis:

Can a Shapley value-based forecast combination outperform an equally weighted forecast combination in volatility forecasting?

To answer the question in this paper, first, the parameters of the models will be estimated in two different environments. First a series of simulations will be done where a data-generating process (DGP) simulates daily returns and realized variances. The second environment is based on a real-life data set of the FTSE100. Via the volatility models, the Shapley values can be calculated to construct the weights in the combination. Finally, out-of-sample data will be used to calculate forecast metrics and conduct tests to evaluate and compare different approaches to weighting.

As mentioned there is a variety of models to forecast volatility. There are simpler models, such as a random walk or also generalized autoregressive conditional heteroscedasticity (GARCH) models which are more sophisticated. An elaborate description and more models will be given in Section 3. These different models raise the question of whether superior models will outweigh simpler models significantly based on their Shapley value. Furthermore, is combining based on these weights beneficial at all? Or would it be better to just use OLS coefficients?

As said, volatility has a big impact on portfolios as it determines the risk of certain assets. Thus, when getting a better forecast this increases the insights on certain decisions of asset allocation, which is useful for asset managers. Furthermore, this method of combining forecasts based on Shapley values is rather new. The application of volatility models is merely an example. If it shows that Shapley-based combinations show promising results, they can be applied in forecasting as a concept that is not bounded by any field. Hence, this research is thus relevant on a scientific basis but also very interesting for practical applications.

Research on volatility presented itself decades ago as well as the innovation of different models and their performances. Being able to make good volatility forecasts is beneficial as it plays a crucial role in the financial world. [Andersen et al., 2005] states that the trade-off between risk and expected return, where risk is associated with some notion of price volatility, constitutes one of the key concepts in modern finance. As such, measuring and forecasting volatility is arguably among the most important pursuit, in empirical asset pricing finance and risk management. It is crucial to understand the time-varying element of volatility (volatility clustering). This plays an important role in asset Value-at-Risk (VaR) and Expected Shortfall (ES). The current literature on this time-varying volatility is sufficient however the aspect of combining these models based on Shapley values does not occur yet. Hence, it can contribute if it shows that it is beneficial to apply this method and give insights into its broader application.

As mentioned, the literature on the concept of volatility is substantial. For example [Andersen et al., 2005], who provide insights on key theoretical developments and also an empirical application. Furthermore, [Brooks and Persand, 2003] has researched how to determine the effectiveness of certain forecasts based on evaluation measures. [Christoffersen and Diebold, 2000] looked at the relevance of volatility forecasting in financial risk management, taking into account what horizon is looked upon.

That same substantial presence in the literature is also there for forecast combinations. Empirical results suggest that a simple equal-weighted average of survey forecasts outperform the best model-based forecasts for a majority of macroeconomic variables and forecast horizons. [Aiolfi et al., 2010] considered multiple types of models and combinations. They conclude that empirical results suggest that an average of survey forecasts outperform the best model-based forecasts for a majority of macroeconomic variables and forecast horizons. [Claeskens et al., 2016] did a theoretical approach on forecast combinations. Comparing equally-weighted cases with 'optimal' combinations and standard models. They state that if the weights are random rather than fixed and are taken into account during the optimal derivation, then that creates a biased combination, and its variance will be larger than in the fixed-weight case. Furthermore, the optimal combination will not automatically outperform the equal-weight case or even the original forecasts. [Smith and Wallis, 2009] explains how simple combinations tend to outperform sophisticated model combinations in empirical examples due to the effect of the finite-sample error in estimating the combining weights. Finally, Shapley values were introduced in 1953 by Lloyd Shapley in game theory [Shapley, 1953]. Shapley values are used in several applications, for example, [Winter, 2002] discusses its applicability in cost allocation, and [Rozemberczki et al., 2022] uses it in a machine learning environment.

[Mishra, 2016](2016) wrote about tackling multicollinearity and the usefulness of Shapley value-based weights. She states that strong multicollinearity harms the confidence intervals of linear regression coefficients. Although it does not affect the R^2 of the regressors or the unbiasedness of the estimated coefficients associated with them. It does, however, inflate their standard error often such that, although R^2's could be very high. Individual coefficients may all have poor Student's t values. Thus, strong multicollinearity may lead to failure in rejecting a false null hypothesis of the ineffectiveness of the regressor variable to the regressand variable. Furthermore, [Franses, 2023] looked at a forecast combination based on Shapley values which came from combinations of R^2's, which will be the basis of this thesis.

Regarding the research question, the results show that in the Monte Carlo simulations, the Shapley approach outperforms the average weights significantly in all the tests and cases. The same applies to the FTSE100 case. This paper has the following structure: Section 2 contains a description of the data and DGP, Section 3 elaborates on the methodology, Section 4 provides the results of the empirical exercises and Section 5 contains a conclusion.

2 Data

This section describes the two data environments used in this research. First the DGPs will be proposed in order to work with Monte Carlo simulations. Secondly, the FTSE100 and its statistical properties will be discussed.

2.1 Monte Carlo Simulation

For the simulation part of this research, simple returns (in %) and realized variances are simulated in a way that matches the habits of real-life data such as volatility clustering. The data consists of 5295 simulated observations mimicking the period 2000-2020 for convenience regarding the real-life data set. The real-life data will be split up in a 60/20/20 ratio and some deviations to check whether the models make sense, determine weights and eventually conduct comparison tests respectively. For simplicity sake, the first 70% of the generated data points will be used to test the models and simultaneously to determine the weights. This because the models already proved to be significant, this way it is computational convenient regarding the Monte Carlo simulations. Furthermore, otherwise it would disregard at least half of the generated data points. The other 30% will be used for out-of-sample testing. The main reason to use a Monte Carlo simulation is to diminish the amount of uncertainties in the results of the tests and loss functions. A table with the distribution of the sample can be found in the Appendix. For the Monte Carlo simulation 100 different series will be simulated. In the calculated metrics the average of 500 runs will be taken, this by randomly selecting one of the simulated series and replace it back into the sample.

Since returns often show no particular autocorrelation, in this simulation, the returns have been done via a Brownian motion. This way all the returns are independent. Furthermore, due to the fatter tails, instead of the normal distribution, it is $T(25)$ distributed. Furthermore, q_t takes an uniform distributed value on a specified interval per time to simulate volatility spikes.

$$r_t = 100\%(\frac{R_t}{R_{t-1}} - 1); \qquad R_t = R_{t-1} + T_t q_t, \qquad T_t \sim \mathcal{T}(25) \tag{1}$$

As mentioned the variable q_t follows a specified uniform distribution per interval. Those are the following:

$$\begin{array}{ll} t \in \{1, 2000\} & q_t \sim U[1; 1.5] \\ t \in \{2001, 2100\} & q_t \sim U[1; 2.5] \\ t \in \{2101, 4000\} & q_t = 1 \\ t \in \{4001, 4100\} & q_t \sim U[1; 3.5] \\ t \in \{4101, 5295\} & q_t = 1 \end{array}$$

The shocks have a minimum/maximum of -/+40%, otherwise extreme outliers would occur and the simulation would not be reliable. When a random shock outside this interval occurs an uniform random value is drawn on the interval [-40,40].

Furthermore, for the realized variances the historical volatility is taken by averaging the last T squared observations. Here T=22 as it represents one month of trading days:

$$RV_t = \sum_{i=1}^{22} r_{t-i}^2 \tag{2}$$

The simulated realized variance will then be converted into annualized volatility by multiplying by 252 trading days and taking the square root as volatility is a standard deviation instead of a variance; ($\text{Vol}_t = \sqrt{252 RV_t}$). As certain assumptions are done about the distribution of r_t, descriptive statistics are computed and reported in Table 1.

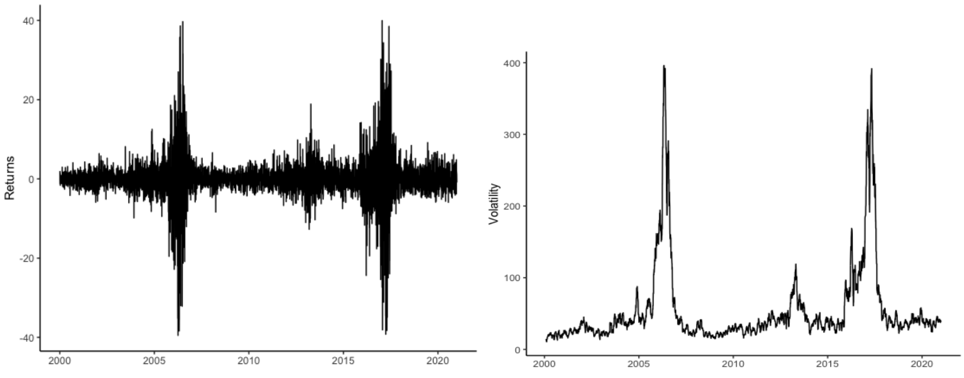
	Mean	SD	Skewness	Kurtosis
r_t	-0.004	4.051	0.096	9.296
Vol.	34.073	54.565	1.857	8.667

Table 1

Descriptive Statistics of simulated returns and volatility. First the values per simulation was calculated, then the average of the 100 series' results was taken.

First, the average return r_t is as expected almost exactly at zero. The standard deviation is slightly above 4 which is way higher than the often assumed standard normal distribution. This however, was to be expected due to the constructed DGP. The normal distribution in general does not apply as well. The skewness is around the expected value of zero however there is quite some excess kurtosis. This leads to having a p-values of 0.000 in the Jarque–Bera test [Jarque and Bera, 1980]. This needs to be taken into account when computing the GARCH models for volatility modeling as it uses maximum likelihood. Further discussion can be found in Section 3.1.4.

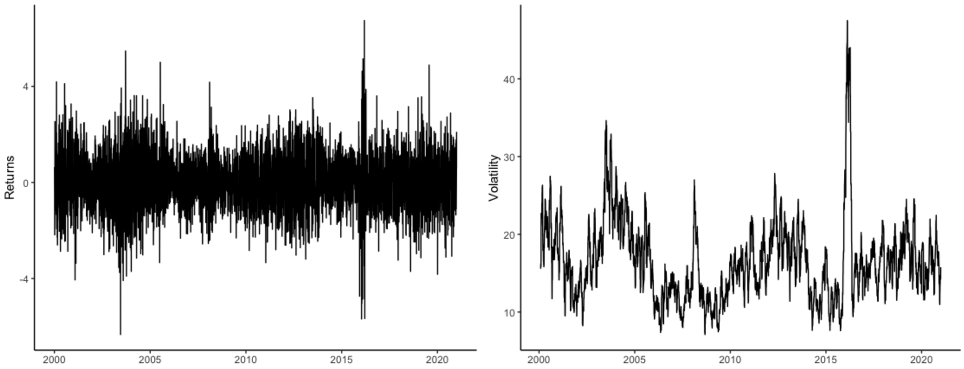
The following figures are two plots of simulated returns and volatilities. For both returns, the series is centered around zero as real-life returns do.



(a) Example 1, Simulated Daily Returns

(b) Example 1, Simulated Annualized Volatility

Figure 1



(a) Example 2, Simulated Daily Returns

(b) Example 2, Simulated Annualized Volatility

Figure 2

Figure 1a-2b: As can be seen in Figure 1a the returns are capped at 40%, this however is not the case in every simulation (Figure 2a). This way, all levels of returns (and thus the realized variance) are taken into account in order to get a successful Monte Carlo simulation who diminishes the uncertainties of all kinds of situations

2.2 FTSE100

As mentioned, the period 2000-2020 is mimicked as the real data set used in this research is from the FTSE100. This is the index of the London Stock Exchange (LSE)[2] consisting of the 100 biggest companies in the UK. It contains the daily returns in percentages and the realized variance (5-minute intra-day returns) as a benchmark for the volatility. The data ranges from January 2000 to December 2020 ($n = 5295$). The data will be split up into three parts. The first part (4/1/2000-31/12/2012) will be used to estimate the models. The second part (2/1/2013-30/12/2016) will be used to estimate the Shapley weights based on the R^2 of all the regressions. The final sample (3/1/2017-31/12/2020) will be used for evaluating the forecast accuracy. These samples give an approximate 60/20/20 partition respectively. Furthermore, to check for robustness, two other samples (50/25/25) and (70/15/15) will be used to check the sensitivity of the models, weights, and parameters.

The four moments of the returns r_t are computed in the three partitions. For all three samples, the average return is around zero. The standard deviations are closer to 1 which would indicate tending towards the standard normal distribution. Often the normal distribution is assumed for stock returns. However, this is again not at issue. The Jarque-Bera statistics again have p-values of 0.000 for all three partitions [Jarque and Bera, 1980]. The skewness in the first two samples is close to zero, however, the kurtosis in all three samples is way above three which means fatter tails and no normal distribution:

[2] https://www.londonstockexchange.com/indices/ftse-100

height	Mean	SD	Skew.	Kurt.
2000-2012	-0.018	1.257	-0.180	8.845
2013-2016	0.018	0.901	-0.133	5.218
2017-2020	0.009	1.086	-1.337	17.108

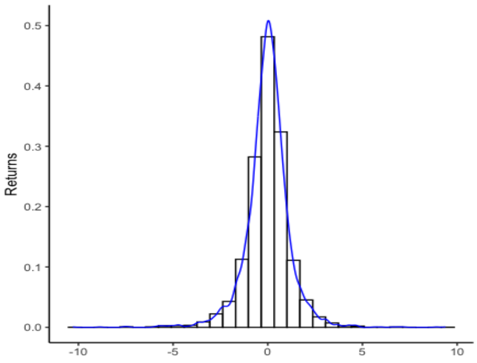
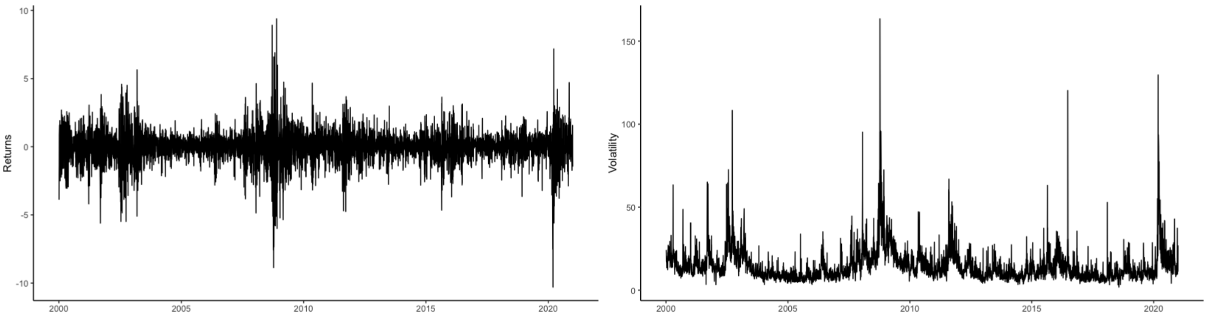


Table 2: Descriptive Statistics returns FTSE100 and histogram of the full sample.

Figure 4a shows the plot of the returns in the FTSE100 2000-2020. The shocks in the plot in 2007 and 2020 are those of the Grand Recession and the Covid-19 crisis respectively. These led to enormous spikes in the volatility which can be seen in Figure 4b.



(a) Daily Returns

(b) Annualized Volatility

Figure 4: Returns and Volatility FTSE100 2000-2020

3 Methodology

In this section, all the procedures, models, and tests will be described that have been done in the research. The main goal is to compare a combination of forecasts where the weights are determined in three different ways.

The most simple weights are those by taking the arithmetic average of the n forecasts available, this will be referred to as $\frac{1}{n}$ weights. The second approach, the main topic of this research, the Shapley weights (SH) which will be discussed in Section 3.2. The third and last type of weight is the weight based on the coefficients of an ordinary least squares (OLS) regression of the volatility on all the forecasts, these will be referred to as coefficient weights (CW).

The first step in this research is to obtain forecast models (Section 3.1). The models are used to first check whether they make sense. Secondly, the different weights as mentioned in the previous paragraph are calculated. Finally, tests are conducted to evaluate the different types of combinations (Section 3.3).

3.1 Models

3.1.1 Realized Volatility

The dependent variable in this research is the annualized realized volatility. The realized volatility is calculated by taking 252 (trading days) times the realized variance and then taking the square root, $\text{Vol}_t = \sqrt{252 RV_t}$. The realized variance is calculated by taking the sum of the squared intra-day returns based on 5 minute intervals in a trading day. A often used proxy for volatility is simply r_t^2 on a given day [Patton, 2011]. The benefit of the realized variance metric opposed to r_t^2 is that the realized variance reflects what actually happened on a trading day. A share price could for example start at 100 an end at 101. The r_t^2 would then imply a volatility of 1. However, the share price could also have dropped to 95 and risen back to 101. This means that there occurred more movement in the share price then the r_t^2 captured.

3.1.2 Random Walk

The first model is a simple method to estimate volatility, the random walk. This is a measure based on the previous value added with a standard normal random variable. In a regular random walk the forecast is simply "calculated" by taking the previous value. In this research however, a standard normal randomly distributed variable w_t is added as well too, furthermore ϵ_t is the usual error term. This gives the following formulation:

$$\hat{\sigma}_t^2 = RV_{t-1} + w_t + \epsilon_t, \qquad w_t, \epsilon_t \sim \mathcal{N}(0,\, 1) \tag{3}$$

Previous literature shows that it is hard in most cases to beat the random walk in volatility forecasting [MCMillan et al., 2000]. Furthermore, it has a big advantage because of its simplicity. The drawback of course is the forecast window being small as its variance increases linearly when forecasting further ahead. For example, y_t follows a standard normal distribution as well as w_t in Formula 3. In that case, the one-step ahead forecast has a variance of $V(y_{t-1}) + V(w_t) + V(\epsilon_t) = 1 + 1 + 1 = 3$. Then the two-step ahead forecast has a variance of $V(\hat{y_t}) + V(w_{t+1}) + V(\epsilon_t) = 3 + 1 + 1 = 5$ etc.

Since volatility cannot be negative but the prediction due to this random walk can, when a negative predicted value occurs this will be re-estimated till it is positive. Furthermore, this research uses a "second-order" random walk in the Monte Carlo simulations. This will be done by taking the estimate in Equation 3, $\sigma_{1,t}^2$, and adding another standard normal randomly distributed variable g_t. Per definition the estimate of $\hat{\sigma}_{2,t}^2$ should in general be worse as g_t increases the uncertainty. Also here applies that when the estimate is negative it will be re-estimated:

$$\hat{\sigma}_{2,t}^2 = \sigma_{1,t}^2 + g_t + \epsilon_t, \qquad f_t, \epsilon_t \sim \mathcal{N}(0,\, 1) \tag{4}$$

3.1.3 Hetero Autoregressive Model

The Heterogeneous Autoregressive Model (HAR) as proposed by Corsi (2009) makes use of lags of previously captured realized volatilities. In this way, it is an autoregressive model using three lags. Namely one day, the average of the previous week (five days), and the average of the previous month (22 days). This constitutes into the following formulations:

$$\begin{aligned}
RV_{t-1}^{d} &= RV_{t-1}\\
RV_{t-1}^{w} &= \frac{1}{5}\sum_{i=1}^{5} RV_{t-i}\\
RV_{t-1}^{m} &= \frac{1}{22}\sum_{i=1}^{22} RV_{t-i}\\
\hat{\sigma}_t^2 &= \beta_0 + \beta_1 RV_{t-1}^{d} + \beta_5 RV_{t-1}^{w} + \beta_{22} RV_{t-1}^{m} + \epsilon_t
\end{aligned} \tag{5}$$

The HAR model is estimated using OLS and is predominately a good fit for certain features of financial data such as long memory and fat tails [Corsi, 2009]. The HAR model allows for an easy estimation, it can simply approximate long memory and is parsimonious. Given these reasons, it is widely used within the research community.

3.1.4 GARCH Models

The GARCH models are often used in volatility forecasting and were invented in 1986 by Bollerslev [Bollerslev, 1986] as an addition to the ARCH models in 1982 [Engle, 1982]. The specification of the GARCH models consists of a return equation and a variance equation. Furthermore, the GARCH parameters are estimated via maximum likelihood.

The first GARCH model is the GARCH(p, q) which includes squared lagged errors and lagged estimated variances. In this research, the order is (1,1) with the following formula:

$$\begin{aligned}
r_t &= \mu + \epsilon_t\\
\hat{\sigma}_t^2 &= \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2\ , \quad (p, q) = (1, 1)
\end{aligned} \tag{6}$$

To start the recursion in the GARCH(1,1) the unconditional variance $E(\sigma^2) = \bar{\sigma}^2$ is used as the first observation . To calculate this unconditional variance there are some assumptions in time series analysis. Namely $E(\sigma_t^2) = \epsilon_t^2$ and $E(\sigma_t^2) = E(\sigma_{t-1}^2)$. This leads to the derivation of $\bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta}$, the derivation can be found in Formula 15 in the Appendix.

The second GARCH model is the component GARCH (C-GARCH). The volatility, which is measured by the conditional variance of stock returns, is decomposed into a long- and short-run component [Lee and Engle, 1993].

$$\begin{aligned}
q_t &= \omega + \rho q_{t-1} + \phi(\epsilon_{t-1}^2 - \sigma_{t-1}^2)\\
\hat{\sigma}_t^2 &= q_t + \alpha(\epsilon_{t-1}^2 - q_{t-1}) + \beta(\sigma_{t-1}^2 - q_{t-1})
\end{aligned} \tag{7}$$

The C-GARCH has two differences compared to the GARCH. First of all, the ω in Formula 6 is constant whereas in Formula 7 it is time-varying. q_t is dependent on its lag and the difference between the squared lagged error and the lagged estimated variance. The second difference is that in the σ_t^2 part of the C-GARCH the ARCH and the GARCH part are reduced with q_{t-1}. The C-GARCH is especially a good fit for sets with fatter tails [Lee and Engle, 1993]. The unconditional variance is $\frac{\omega}{1-\rho}$ (Appendix Formulas 16&17).

The final GARCH model considered is the Glosten, Jagannathan, and Runkle-GARCH(1,1) (gjr-GARCH(1,1)) with the following formulation:

$$\hat{\sigma}_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \gamma\epsilon_{t-1}^2 I[\epsilon_{t-1} \leq 0] + \beta\sigma_{t-1}^2 \tag{8}$$

The difference is in the error term which now contains an indicator function making a distinction between positive and negative errors. Negative news tends to have a bigger influence on volatility than positive news, so including this leverage term can capture this bigger effect [Glosten et al., 1993]. The theoretical expected value of $I[\epsilon_{t-1} \leq 0] = 0.5$. The derivation of the unconditional variance can be found in the Appendix Formula 18.

GARCH models are estimated using maximum likelihood (ML). The ϵ_t term is used for this and its distribution therefore is necessary. For this, often a normal distribution is assumed however in practice this does not always apply. In this research, in the maximum likelihood the t-distributed is assumed due to the excess kurtosis (Table 1).

3.2 Shapley Values

In order to evaluate the relative contribution of volatility forecasts, a series of regression models is estimated. The R^2 is used to decompose the relative contributions of each forecast method as an explanatory variable of the realized volatility. This, by evaluating the R^2 for all possible combinations of explanatory variables and taking a weighted average of each variables' contribution to the R^2. This constitutes into Shapley values. Initially these values where used to calculate the utility of players in cooperative game theory [Shapley, 1953]. Its concept has found applications in fields such as economics and mathematics. Furthermore, Shapley values are convenient to use as they are easy to interpret.

The main purpose of this research is to check whether a combination based on Shapley values outperforms $\frac{1}{n}$ weights in its forecasting ability. The Shapley weights are calculated in the following way. First, the different forecasts have been made to obtain these as explanatory variables $X_1, ..., X_K$, in this research K=5. Then separate regressions are performed to obtain the R^2's of the regressions. All the combinations of one explanatory variable $\binom{5}{1}$ first, then all the combinations of two explanatory variables $\binom{5}{2}$, etc. This gives 5+10+10+5+1=31 different R^2's. In total, there are $2^K - 1$ regressions performed, where K is the amount of explanatory variables [Franses, 2023]. The R^2's are used to find the individual contribution of an explanatory variable X_i. This is done via the following formula [Chantreuil and Trannoy, 1999]:

$$\begin{aligned} SH_j &= \sum_j \frac{(s-j)!(k-s)!}{k!}(R^2(S) - R^2(S\{j\})) \\ s_j &= \frac{SH_j}{R^2_{12345}} \end{aligned} \tag{9}$$

In this formula, SH_j is the net contribution of variable j before scaling it to $R^2_{12..k}$, which is the R^2 of the regression with all the variables. K is the amount of variables and S is a subset of K containing $|S|$ explanatory varables. An example for $K = 4$ can be found in the Appendix of Franses 2023. After obtaining the Shapley weights the following linear combination will be our Shapley-based forecast:

$$y_t = \sum_{j}^{K} s_j f_{j,t} + \epsilon_t \tag{10}$$

In this formula f_j is an individual forecast. This combination, y_t, will be compared to two different forecast combinations, the $\frac{1}{n}$ weights and the coefficient weights. The coefficient weights will be normalized by dividing each weight with the sum of the weights. This, as their sum is not necessarily equal to one which is the case for the Shapley weights.

3.3 Evaluating Forecasts

When the models are estimated and the proper weights have been calculated, the next step is to test out-of-sample how the models perform. As the goal is to evaluate the combinations rather than the models themselves, new model parameters are estimated for the out-of-sample data. However, the combination weights are based on the second sample. Thus, it is still valid to evaluate it as an out-of-sample. For this evaluation, several metrics have been proposed and used in the field of forecasting. For this research the following five metrics are chosen [Andersen et al., 2005]:

$$\begin{aligned}
MAD &= N^{-1}\sum_{i=1}^{N}|y_i - f_i| \\
ME &= N^{-1}\sum_{i=1}^{N}(y_i - f_i) \\
RMSE &= \sqrt{N^{-1}\sum_{i=1}^{N}(y_i - f_i)^2} \\
R^2Log &= N^{-1}\sum_{i=1}^{N}[log(y_i^2 f_i^{-2})]^2 \\
QLike &= N^{-1}\sum_{i=1}^{N}(log(f_i^2) + y_i^2 f_i^{-2})
\end{aligned} \tag{11}$$

Here, y_i is the actual value and f_i is the forecast. For all the five metrics hold, the closer they are to zero the better the result. The Mean Absolute Deviance (MAD) calculates on average how far a forecast is from the actual value. The Mean Error (ME) is the only value that can get negative, as it simply takes the mean of the actual values and subtracts the mean of the forecasts to check for a certain bias. If the ME is positive the forecast underestimates and vice versa. The Root Mean Squared Error (RMSE) is quite similar to the MAD although it penalizes bigger errors by taking the square. Then the R^2Log which takes the average of the squared Log of the squared ratio $\frac{y_i}{f_i}$. Finally, the QLike which consists of the sum of the Log squared forecasts and again uses the squared ratio $\frac{y_i}{f_i}$. [Patton, 2011] further discusses these loss functions in a volatility forecast environment.

Not only the metrics above are useful in order to determine what a good forecast can be. In forecasting there is a trade off between a point forecast and the variance it comes with. A forecast can be good compared to another however its confidence intervals could be substantially bigger. This could lead to meaningless forecasts. Since the combinations provided in this research are linear combinations its variance is simply calculated by the following formula, here w_i is a weight and X_i an explanatory variable.

$$Var(\sum_{i=1}^{N} w_i X_i) = \sum_{i=1}^{N} w_i^2 Var(X_i, X_j) + \sum_{1 \leq i < j \leq N} 2 w_i w_j Cov(X_i, X_j) \tag{12}$$

This is computed using a weight vector w and the variance matrix V. Formula 12 can be reformulated as $w^T V w$. The purpose of the calculation is to see whether besides the outperformance on point forecasts, the confidence intervals also are useful and to see whether or not there is a trade-off.

Another method to evaluate forecasts is the Mincer Zarnowitz regression [Mincer and Zarnowitz, 1969]. This regression simply regresses the actual value of the forecast in the out-of-sample set. It has the following formulation:

$$y_i = \alpha + \beta f_i + \epsilon_i \tag{13}$$

If a forecast is accurate, the ideal situation is where α=0 and $\beta = 1$. This, assuming that $E(\epsilon_i) = 0$, would then come down to $y_i = f_i$ which is the target of our forecast. Furthermore, a high R^2 is expected as the explanatory variable is already a forecast of y_i.

Finally, a Diebold Mariano test can be done to find if one forecast is significantly outperforming the other [Diebold and Mariano, 2002]. The test uses a variable d_t, where $d_t = e_{1,t}^2 - e_{2,t}^2$. Meaning that it takes the difference between the squared errors of two different forecasts. The test statistic of the Diebold Mariano test is

$$DM = \frac{\bar{d}}{\sqrt{Var(d)}} \sim \mathcal{N}(0,\, 1) \tag{14}$$

When the DM test statistic is negative it means that the first forecast outperforms the second one and vice versa (if the corresponding p-value is below the chosen α).

4 Results

The result section be split up in the Monte Carlo simulation section and in the section about the FTSE100. First the weights will be discussed, than the forecast metrics and finally the Mincer Zarnowitz regression and the Diebold Mariano test. Also the model specifications of the FTSE100 will be presented.

4.1 Monte Carlo Simulation

4.1.1 Shapley Weights

The first step after simulating the 100 series of returns and realized variances is to estimate the models as described in Section 3. One remark, for the HAR model, is that the first autoregressive term β_1 each time is multiplied by a uniformly distributed random variable between 0.5 and 1.5 as otherwise the forecast will be to good due to the DGP. The forecasts form a matrix which can be seen as the explanatory variables who's Shapley weights can be computed. The simulated series are chosen randomly 500 times with replacement. From where the average of these computed scores are taken. The following table provides the average weights of the Shapley approach.

	RW	RW_2	GARCH(1,1)	gjr-GARCH	HAR
Shapley	0.120	0.102	0.296	0.229	0.252
CW*	0.166	0.026	0.529	0.088	0.191

*Weights normalized to sum up to 1

Table 3: Average Shapley and OLS weights

Table 3 shows the average weights based on the simulations. The GARCH(1,1) is in both approaches the main driver for the forecast combination. In the Shapley approach its weight is half of the CW approach, this again due to the correlation where Shapley values account for. The random walks show different behaviour across the two approaches. In the Shapley weights, there is no big difference between RW and RW_2 whereas in the CW RW_2 is approximately a sixth of RW. This again shows that OLS coefficients almost exclude worse forecasts such as RW_2 while the Shapley approach looks at its individual relative contribution. In order to compute the standard deviation of the forecast combination covariance\correlation matrix of the out-of-sample data is given:

$$\hat{V}\backslash\hat{\rho} = \left(\begin{array}{cccccc} & RW & RW_2 & GARCH(1,1) & gjr-GARCH & HAR \\ RW & 30.3 & 0.52 & 0.10 & 0.02 & 0.08 \\ RW_2 & 18.7 & 39.3 & 0.13 & 0.04 & 0.03 \\ GARCH(1,1) & 5.8 & 5.6 & 15.8 & 0.49 & 0.23 \\ gjr-GARCH & 4.49 & 4.6 & 9.1 & 9.3 & 0.02 \\ HAR & 4.3 & 4.1 & 7.0 & 5.5 & 55.9 \end{array}\right)$$

Note: the bottom left matrix including the diagonal can be read as the (co)variance matrix $\hat{V}$. The upper right matrix as the correlation matrix $\hat{\rho}$.

Using Formula 12, the standard deviations can be calculated by taking the square root of the computed variance. The Shapley weight approach has the lowest standard deviation with 3.37 followed by the $\frac{1}{n}$ approach ($\sigma = 3.40$) and the coefficient weights are the highest with a standard deviation of 3.67. Furthermore, Figure 5 displays the average Shapley weights of the previous simulations. After approximately 100 runs the Shapley weights start to convert and after 300 there is no real change in the weights.

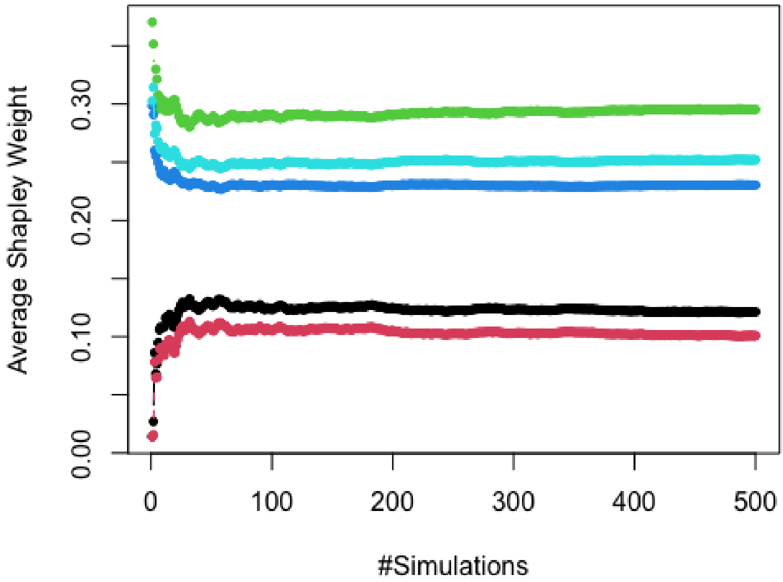


Figure 5: The average Shapley weights of the past n calculated weights of the randomly chosen simulation. From top to bottom: GARCH(1,1), HAR, gjr-GARCH, RW, RW_2. After 100 iterations the average weights become stable, after approximately 300 iterations the average weights do not change.

4.1.2 Evaluation Metrics

	SH	$\frac{1}{n}$	CW
MAD	3.19	3.81	**2.69**
ME	1.25	0.53	**0.21**
RMSE	4.67	5.29	**4.00**
R^2Log	0.07	0.19	**0.06**
QLike	7.35	7.40	**7.35**

Table 4: Average forecast performance metrics of the three weight approaches

Table 4 shows the average forecast metric values of the simulations. The coefficient weights outperform the other two approaches in all the five metrics meaning it has the best out-of-sample performance. The Shapley approach is second to best beating the $\frac{1}{n}$ approach in all the metrics except for the mean error. The Shapley approach compared to $\frac{1}{n}$ tends to have a overall lower deviance (MAD) however a bigger bias (ME). This, looking at the RMSE, comes from bigger outliers in the $\frac{1}{n}$ approach.

4.1.3 Mincer Zarnowitz & Diebold Mariano

First, a Mincer Zarnowitz regression will be done on the simulations. The average of the α, β and R^2 are calculated. Starting with the $\frac{1}{n}$, with -0.54, 0.73 and 0.57 respectively . Then the Shapley weights with on average values of -1.57, 0.85, 0.670. Finally, the coefficient weight approach with α =-1.68, β = 0.84 and a R^2 of 0.678. Comparing the three, the CW approach shows to have the highest R^2 on average. However, its corresponding α and β are further away from the ideal values than the other two approaches. Again all the α's are negative which coincides with the ME in Table 4 as the models overestimate the volatility.

Finally, the average Diebold Mariano statistics are computed of the three different comparison possibilities. The Shapley approach statistically outperforms the $\frac{1}{n}$ with a DM of -12.43 (on average). The CW approach however significantly outperforms the Shapley and $\frac{1}{n}$ approaches with DM-statistics of (-)5.37 and 12.03 respectively. The corresponding p-values to the average DM-statistic are all 0.000. This is also in line with the values in Table 4.

	SH-$\frac{1}{n}$	SH-CW	CW-$\frac{1}{n}$
DM	-12.43	5.37	12.03

Table 5: Average Diebold Mariano statistics for the simulated series

4.2 FTSE100

4.2.1 Models and Weights

The simulated experiment shows promising results to try and use it on a real data set. For this, as mentioned in Section 2.2 the 100 biggest UK-listed companies (FTSE100) are used to create volatility models and calculate Shapley values to determine the weights. Again the procedure of calculating the models first gives the following formulations for the 60/20/20 partition:

Model specifications 60/20/20

		μ	Model
2000-2012	HAR	-	$\hat{\sigma}_t^2 = 0.215 + 0.149RV^d_{t-1} + 0.481RV^w_{t-1} + 0.218RV^m_{t-1}$
	GARCH	0.033	$\hat{\sigma}_t^2 = 0.012 + 0.101\epsilon^2_{t-1} + 0.893\sigma^2_{t-1}$
	C-GARCH	0.034	$\hat{\sigma}_t^2 = q_t + 0.067(\epsilon^2_{t-1} - q_{t-1}) + 0.891(\sigma^2_{t-1} - q_{t-1})$
	q_t (C-GARCH)	-	$q_t = 0.004 + 0.998q_{t-1} + 0.051(\epsilon^2_{t-1} - \sigma^2_{t-1})$
	gjr-GARCH	0.000	$\hat{\sigma}_t^2 = 0.016 + 0.163\epsilon^2_{t-1}I[\epsilon_{t-1} \leq 0] + 0.904\sigma^2_{t-1}$
2013-2016	HAR	-	$\hat{\sigma}_t^2 = 0.411 + 0.028RV^d_{t-1} + 0.276RV^w_{t-1} + 0.164RV^m_{t-1}$
	GARCH	0.032	$\hat{\sigma}_t^2 = 0.050 + 0.163\epsilon^2_{t-1} + 0.782\sigma^2_{t-1}$
	C-GARCH	0.032	$\hat{\sigma}_t^2 = q_t + 0.145(\epsilon^2_{t-1} - q_{t-1}) + 0.751(\sigma^2_{t-1} - q_{t-1})$
	q_t (C-GARCH)	-	$q_t = 0.005 + 0.994q_{t-1} + 0.020(\epsilon^2_{t-1} - \sigma^2_{t-1})$
	gjr-GARCH	0.004	$\hat{\sigma}_t^2 = 0.039 + 0.283\epsilon^2_{t-1}I[\epsilon_{t-1} \leq 0] + 0.815\sigma^2_{t-1}$
2017-2020	HAR	-	$\hat{\sigma}_t^2 = 0.186 + 0.026RV^d_{t-1} + 0.927RV^w_{t-1} - 0.125RV^m_{t-1}$
	GARCH	0.049	$\hat{\sigma}_t^2 = 0.018 + 0.107\epsilon^2_{t-1} + 0.879\sigma^2_{t-1}$
	C-GARCH	0.052	$\hat{\sigma}_t^2 = q_t + 0.100(\epsilon^2_{t-1} - q_{t-1}) + 0.704(\sigma^2_{t-1} - q_{t-1})$
	q_t (C-GARCH)	-	$q_t = 0.005 + 0.992q_{t-1} + 0.036(\epsilon^2_{t-1} - \sigma^2_{t-1})$
	gjr-GARCH	0.034	$\hat{\sigma}_t^2 = 0.018 + 0.121\epsilon^2_{t-1}I[\epsilon_{t-1} \leq 0] + 0.895\sigma^2_{t-1}$

First, the HAR model shows some differences in the parameters for the different periods. The constant starts at approximately 0.2, jumping to 0.4 and then back again to 0.2. As in Table 3 can be seen, the second sub-sample is less volatile thus the constant has a bigger impact on the forecast as the parameters of the autoregressive tend to decline. In the third HAR equation, the weekly average has a big impact looking at the relative values of the parameters β_1, β_5, and β_{22}. This period is relatively volatile, meaning that a weekly average tends to be a good estimator as the volatility is high in a period. For the GARCH models, the μ is rather stable across the periods with some minor fluctuations. The α (or γ in the gjr-GARCH) increases in the 2013-2016 period as the GARCH models show to emphasize more on occurring shocks in the stabler periods. In the C-GARCH model, the q_t equation shows no major shifts in parameters.

After implementing the models, as in the simulation, the 31 regressions per partition are done to receive the R^2's to calculate the Shapley weights. Furthermore, the weights based on the coefficients of the regular regression on all the explanatory are computed.

		RW	HAR	GARCH(1,1)	C-GARCH	gjr-GARCH
50/25/25	Shapley	0.063	0.264	0.199	0.203	0.272
	CW*	0.015	0.614	-1.273	0.979	0.664
60/20/20	Shapley	0.053	0.148	0.226	0.243	0.330
	CW*	0.044	0.245	-1.160	1.215	0.657
70/15/15	Shapley	0.051	0.175	0.218	0.237	0.319
	CW*	0.012	0.343	-1.050	1.050	0.645

*Weights normalized to sum up to 1

Table 7: Weights of models for the three different partitions

First, comparing the weights across the periods. For the Shapley weights, the random walks weight fluctuates around 5%. This, as it does not perform well but still has some forecasting power. The HAR model does well in the first sample having the second highest weight, although declining in the second and third models where it is second to last. Here the difference can be seen comparing the DGP to the real-life set. The GARCH models take most of the weight ranging from around 65% in the first period to almost 80% in the second period. For the coefficient weights the random walk is, except for the second period, almost at zero showing its randomness as in a normal regression it would not have very strong explanatory power. The HAR model, again has a relatively high weight in the first sample partition but decreases in the other two. The gjr-GARCH model remains to hold a stable weight of around 0.650 over the three periods. Interesting to see is the GARCH and C-GARCH. Their respective correlations in the three periods are 0.994, 0.991, and 0.981. Due to this high correlation, the weights look to be chosen to almost diversify the combination. This way the two models almost cancel each others' contribution out. This happens in all three periods; in the first sample the GARCH takes a more negative value as in this period the GARCH overestimates the volatility slightly more than the C-GARCH with 1.83 and 1.77 percentage points on average respectively. In the second sample distribution the two weights are almost the same and in the third they are the same where the weights are practically canceling each other out. This does not necessarily mean that the individual models do not contribute at all due too to the forecast not having the same values. However, given the track record of the GARCH(1,1) it does not make sense to use this combination approach. This shows the relevance of the forecast combination based on Shapley values.

Note that as the coefficient weights can be negative and above 1/ below -1), the variance of an individual regressor ($\mathrm{E}(w_i^2 X_i^2)$) can be way higher. For example comparing the GARCH(1,1) in the 60/20/20, the Shapley weight squared is 0.05 and that from the CW approach is 1.36. This can lead to very high standard deviations when the variance of the corresponding explanatory variable is high. Which eventually can lead to higher and less reliable confidence intervals. This makes methods such as OLS to determine weights less convenient. However, these negative weights do compensate when the covariance is also high which is the case with the GARCH(1,1) and the C-GARCH. When the standard deviations are calculated on the three different partitions for the three methods, the coefficient based weights show the lowest variance using the out-of-sample data. The Shapley and $\frac{1}{n}$ approach alternate per period with the second lowest variance. All in all, the standard deviations do not differ that much (see Table 8).

σ	50/25/25	60/20/20	70/15/15
SH	7.7	8.2	8.7
$\frac{1}{n}$	7.6	8.2	9.0
CW	**7.6**	**8.1**	**8.0**

Table 8: Standard deviations per weight approach per partition

4.2.2 Evaluation Metrics

Again, for the different samples the metrics mentioned in Section 3.3 are computed, the bold figures are the lowest across the three methods:

	50/25/25			60/20/20			70/15/15		
	SH	$\frac{1}{n}$	CW	SH	$\frac{1}{n}$	CW	SH	$\frac{1}{n}$	CW
MAD	4.20	4.44	**3.84**	3.88	4.20	**3.41**	4.23	4.26	**3.79**
ME	-1.87	-2.15	**-1.37**	-1.46	-1.98	**-0.08**	-1.49	-1.22	**-0.33**
RMSE	6.70	6.86	**6.51**	6.18	6.28	**5.87**	6.67	6.74	**6.56**
R^2LOG	0.59	0.65	**0.51**	0.54	0.64	**0.45**	0.55	0.56	**0.45**
QLIKE	**6.14**	6.16	6.14	**6.03**	6.05	6.03	6.21	6.23	**6.20**

Table 9: Forecast metrics for three sample partitions

Compared to the simulation in Table 4 the Shapley approach is again not the best performing approach being beaten by the CW on almost every metric for all the three sample distributions. The Shapley approach only performs the best on the QLIKE in two of the three cases. The $\frac{1}{n}$ is again the worse method, being beaten by the Shapley approach for all the three partitions. Across the three sample distributions, the values maintain approximately the same indicating no sensitivity in the outcome when adapting certain features or parameters which is beneficial from a robustness perspective. The 60/20/20 partition of the observations has the lowest values in Table 9 meaning that this, among the three, is the most favorable.

In the simulation, the Shapley weights showed the best results and in the FTSE100 data, the coefficient weights proved to be the best. This could be due to the presence of multicollinearity, meaning that an explanatory variable is (too much) correlated to other explanatory variables. Therefore the variance inflation factor (VIF) is calculated. This is done by regressing the explanatory variable X_i on all the other variables. Then the VIF is $\frac{1}{1-R^2}$ where the R^2 of the regression is used. A VIF of 1 indicates no correlation between the other variables, a VIF of 1 to 5 indicates moderate correlation, and above 5 severe correlation.

Calculating the VIFs, the GARCH and C-GARCH stand out with 76.2 and 67.6 respectively, indicating severe correlation. The problem in this instance is that for a normal OLS regression, some correlation between certain variables occurs. In this case, it makes sense that the forecasts are correlated as they all individually are already prediction of the same variable. However, when this correlation is too high (in this example the GARCH and C-GARCH) one should be removed and replaced with a different, less correlated, volatility forecaster in the case of OLS.

Furthermore, again the $\frac{1}{n}$ shows to be inferior to the Shapley approach having less favorable values for all the three partitions of the data set. This indicated, that not only in the simulations but also in real-life it is more useful to determine weights based on Shapley values instead of the arithmetic average.

4.2.3 Mincer Zarnowitz & Diebold Mariano

	50/25/25			60/20/20			70/15/15		
	SH	$\frac{1}{n}$	CW	SH	$\frac{1}{n}$	CW	SH	$\frac{1}{n}$	CW
α	-1.27	-1.42	-1.08	-1.29	-1.99	-0.22	-1.81	-0.83	-1.93
β	0.96	0.95	0.98	0.99	1.00	1.01	1.02	0.974	1.11
R^2	0.567	0.556	0.574	0.648	0.652	0.663	0.649	0.637	0.651

Table 10: Mincer Zarnowitz regression statistics for the three sample partitions

In the Mincer Zarnowitz regressions, all three partitions show similar results. The α's are fluctuating between -2 and -0.2, this means that the models in general overestimate the volatility. The β's are all very close to 1.0 which is the ideal result, only for the coefficient weight approach in the last sample the β deviates more than 0.05 from this value. The R^2's range from 0.55 in the first partition to 0.65 in the other two. In the 60/20/20 split, the values are the best having the highest R^2 and the α's being almost exactly 1.0 for all the three methods. This again shows that among the three splits this is the best performing one.

	50/25/25			60/20/20			70/15/15		
	SH-$\frac{1}{n}$	SH-CW	CW-$\frac{1}{n}$	SH-$\frac{1}{n}$	SH-CW	CW-$\frac{1}{n}$	SH-$\frac{1}{n}$	SH-CW	CW-$\frac{1}{n}$
DM	-3.884	2.773	-4.990	-2.352	2.951	-4.260	-1.034	1.219	-1.536
p	0.000***	0.006**	0.000***	0.019*	0.003**	0.000***	0.302	0.223	0.125

Table 11: Diebold Mariano test statistics and p-values

Finally, the Diebold Mariano tests. Across all three partitions, there is no shift in the sign of the DM statistic. The Shapley outperforms the $\frac{1}{n}$ weights and the coefficient weights outperform the Shapley weights. This results is also in line with the findings in Table 9. Furthermore, the significance in the first two samples also remains at the same level around 0.000. However, in the final sample (70/15/15) the significance drops heavily with p-values increasing way above the 5% level. This is due to the out-of-sample set containing only one major shock namely the Covid-19 crisis, whereas the other out-of-sample sets had more shocks. In these samples, the superior approaches had enough data points to significantly outperform the inferior. This amplifies the relevance of robustness checks.

5 Conclusion

This paper investigated combinations of several volatility forecasts based on different weight combinations. In particular, three different combinations were chosen and compared. These are equally weighted, based on Shapley values of the R^2 and regular OLS coefficients. The main question this research tried to answer is:

Can a Shapley value-based forecast combination outperform an equally weighted forecast combination in volatility forecasting?

This research was done using two different environments; the first one used 100 different Monte Carlo simulated returns and realized variances with a data generating processes, 500 runs have been done with the replacement of drawn series. Furthermore, a real-life data set of the FTSE100 has been used for the period 2000-2020. The procedure in short is that the data was split up into three partitions. The first is to check the validity of the models, the second is to estimate the weights of the forecast models, and the third is to test out-of-sample. For the volatility forecasts, several models were chosen. The models used in this research are the following: Random walk, a second-order random walk, GARCH(1,1), gjr-GARCH, C-GARCH, and HAR models.

The Shapley weights showed to take the correlation between the forecasts into account. In the Monte Carlo simulation, the weights of the more advanced regressors showed to be reduced due to the Shapley values looking at individual contribution. For the Monte Carlo simulation, the GARCH(1,1) had the highest weight (29.6%), and in the FTSE100 the gjr-GARCH had the highest weight (33.0%) which only took negative shocks into account. These weights also contribute to the variance in the linear combination. In the Monte Carlo simulation, the Shapley approach had the lowest standard deviation of 3.37. For the FTSE100, the standard deviations were rather similar however the coefficient weights showed to have the lowest in the three different splits. Furthermore, loss functions were calculated where the forecasts were compared to the actual values using out-of-sample data. In the Monte Carlo simulation, the Shapley approach showed to beat the $\frac{1}{n}$ on four of the five calculated metrics. The coefficient weights showed to beat both the Shapley and $\frac{1}{n}$ approach. A quite similar result occurred in the FTSE100 case. The Shapley approach again is superior to the $\frac{1}{n}$ weights however being inferior to the coefficient weights. Finally, a series of Diebold Mariano tests were done. In the simulation, the average of values was computed (-12.43) which indicates significant outperformance of the Shapley approach versus the $\frac{1}{n}$ benchmark. In the FTSE100 three different splits were looked upon. In all three splits the Shapley approach outperforms the $\frac{1}{n}$ approach, only in the last split it was not significant due to the data and number of observations. This all, concludes that it is beneficial to use Shapley values based on the R^2 to determine weights in combining volatility forecasts rather than the arithmetic average. Shapley weights show a slightly lower standard deviation and do better regarding loss functions. Furthermore, compared to OLS, Shapley includes the forecast model in a manner where the correlation is taken into account, whereas the OLS is not prone to multicollinearity. Finally, the Diebold Mariano test also shows that the Shapley approach significantly outperforms the $\frac{1}{n}$ approach. Still, the performance of the forecast model based on the normalized OLS coefficients tends to outperform the Shapley weight approach. Therefore a recommendation for further research would be the trade-off between Shapley weights and coefficient weights, predominately regarding the multicollinearity. Furthermore, in this research simple returns were used. In a comparison, Log returns could also be applied to see whether this would change results.

References

[Aiolfi et al., 2010] Aiolfi, M., Capistrán, C., and Timmermann, A. (2010). Forecast combinations. *CREATES research paper*, (2010-21).

[Andersen et al., 2005] Andersen, T., Bollerslev, T., Christoffersen, P., and Diebold, F. (2005). Volatility forecasting. NBER Working Papers 11188, National Bureau of Economic Research, Inc.

[Bollerslev, 1986] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

[Brooks and Persand, 2003] Brooks, C. and Persand, G. (2003). Volatility forecasting for risk management. *Journal of Forecasting*, 22(1):1–22.

[Chantreuil and Trannoy, 1999] Chantreuil, F. and Trannoy, A. (1999). Inequality Decomposition Values: the Trade-Off Between Marginality and Consistency. Papers 99-24, Paris X - Nanterre, U.F.R. de Sc. Ec. Gest. Maths Infor.

[Christoffersen and Diebold, 2000] Christoffersen, P. F. and Diebold, F. X. (2000). How Relevant is Volatility Forecasting for Financial Risk Management? *The Review of Economics and Statistics*, 82(1):12–22.

[Claeskens et al., 2016] Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

[Corsi, 2009] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

[Diebold and Mariano, 2002] Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144.

[Engle, 1982] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

[Franses, 2023] Franses, P. H. (2023). Shapley-values based forecast combination.

[Glosten et al., 1993] Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801.

[Jarque and Bera, 1980] Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255–259.

[Lee and Engle, 1993] Lee, G. G. and Engle, R. F. (1993). A permanent and transitory component model of stock return volatility. *Available at SSRN 5848*.

[MCMillan et al., 2000] MCMillan, D., Speight, A., and Apgwilym, O. (2000). Forecasting uk stock market volatility. *Applied Financial Economics*, 10(4):435–448.

[Mincer and Zarnowitz, 1969] Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 3–46. NBER.

[Mishra, 2016] Mishra, S. K. (2016). Shapley value regression and the resolution of multicollinearity. *Available at SSRN 2797224*.

[Patton, 2011] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

[Rozemberczki et al., 2022] Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. (2022). The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*.

[Shapley, 1953] Shapley, L. S. (1953). *17. A Value for n-Person Games*, pages 307–317. Princeton University Press, Princeton.

[Smith and Wallis, 2009] Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

[Winter, 2002] Winter, E. (2002). Chapter 53 the shapley value. volume 3 of *Handbook of Game Theory with Economic Applications*, pages 2025–2054. Elsevier.

A Appendix

A.1 Tables and derivations

	50/25/25	60/20/20	70/15/15	Monte Carlo
Model validation	23-2648 ($n = 2626$)	23-3273 ($n = 3251$)	23-3707 ($n = 3685$)	23-3723 ($n = 3701$)
Weight calculation	2649-3972 ($n = 1324$)	3274-4285 ($n = 1012$)	3708-4500 ($n = 793$)	23-3723 ($n = 3701$)
Out-of-sample	3973-5295($n = 1323$)	4286-5295($n = 1010$)	4501-5295($n = 795$)	3274-5295 ($n = 2022$)

$$\begin{aligned} E(\sigma_t^2) &= E(\omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2) \\ &= \omega + \alpha E(\epsilon_{t-1}^2) + \beta E(\sigma_{t-1}^2) \\ &= \omega + \alpha E(\sigma_t^2) + \beta E(\sigma_t^2) \\ \bar{\sigma}^2 &= \omega + \alpha\bar{\sigma}^2 + \beta\bar{\sigma}^2 \\ \bar{\sigma}^2 &= \frac{\omega}{1-\alpha-\beta} \end{aligned} \tag{15}$$

$$\begin{aligned} E(q_t) &= E(\omega + \rho q_{t-1} + \phi(\epsilon_{t-1}^2 - \sigma_{t-1}^2)) \\ &= \omega + \rho E(q_{t-1}) + \phi E(\epsilon_{t-1}^2 - \sigma_{t-1}^2) \\ &= \omega + \rho E(q_t) + \phi[E(\epsilon_{t-1}^2) - E(\sigma_{t-1}^2)] \\ &= \omega + \rho E(q_t) + \phi[E(\epsilon_{t-1}^2) - E(\epsilon_{t-1}^2)] \\ E(q_t) &= \omega + \rho E(q_t) \\ E(q_t) &= \frac{\omega}{1-\rho} \end{aligned} \tag{16}$$

$$\begin{aligned} E(\sigma_t^2) &= E(q_t + \alpha(\epsilon_{t-1}^2 - q_{t-1}) + \beta(\sigma_{t-1}^2 - q_{t-1})) \\ &= E(q_t + \alpha\epsilon_{t-1}^2 - \alpha q_{t-1} + \beta\sigma_{t-1}^2 - \beta q_{t-1}) \\ &= E(q_t) + \alpha E(\epsilon_{t-1}^2) - \alpha E(q_{t-1}) + \beta E(\sigma_{t-1}^2) - \beta E(q_{t-1}) \\ \bar{\sigma}^2 &= \frac{\omega}{1-\rho} - \alpha\frac{\omega}{1-\rho} - \beta\frac{\omega}{1-\rho} + \alpha\bar{\sigma}^2 + \beta\bar{\sigma}^2 \\ \bar{\sigma}^2 &= \frac{\frac{\omega}{1-\rho}(1-\alpha-\beta)}{1-\alpha-\beta} \\ \bar{\sigma}^2 &= \frac{\omega}{1-\rho} \end{aligned} \tag{17}$$

$$\begin{aligned} E(\sigma_t^2) &= E(\omega + \alpha\epsilon_{t-1}^2 + \gamma\epsilon_{t-1}^2 I[\epsilon_{t-1} \leq 0] + \beta\sigma_{t-1}^2) \\ E(\sigma_t^2) &= \omega + \alpha E(\epsilon_{t-1}^2) + \gamma E(\epsilon_{t-1}^2 I[\epsilon_{t-1} \leq 0]) + \beta E(\sigma_{t-1}^2) \\ E(\sigma_t^2) &= \omega + \alpha E(\sigma_t^2) + 0.5\gamma E(\sigma_t^2) + \beta E(\sigma_t^2) \\ \bar{\sigma}^2 &= \omega + \alpha\bar{\sigma}^2 + 0.5\gamma\bar{\sigma}^2 + \beta\bar{\sigma}^2 \\ \bar{\sigma}^2 &= \frac{\omega}{1-\alpha-0.5\gamma-\beta} \end{aligned} \tag{18}$$

A.2 Code Description

In this research, R-studio and Excel have been used. A brief description will be given about all the steps taken.

A.2.1 Excel

For the FTSE100 data, the models have first been made in R-studio which will be discussed further on. For the three partitions two files are created in Excel, the first one with the estimated models and the second one where the estimated values are converted to annual volatility.

A.2.2 R-Studio

In R-studio the following packages are used; *rugarch*, *tseries*, *ggplot2*, *car*, *lgarch*, *forecast*, *zoo*, *HARModel* and *moments*. Three files are used for programming. The first one specifies and estimates the models for the simulations and the FTSE100, furthermore, the VIF calculation is also in this file. Secondly, the Monte Carlo simulation, The returns, and realized variances for the 100 series are generated following the DGP in Section 2. Then the models are estimated and new matrices are made to fill the model per simulation (for example, all the 100 GARCH(1,1) models are in one matrix of 5295x100). Then in a for loop, a random number i is drawn between the simulations (with replacement). The weight data and out-of-sample data matrices are made by appending the i-th column of the matrices containing the volatility and its corresponding forecasts. All the R^2's of the regressions are then computed and using Formula 9 the Shapley weights are calculated. Finally, a matrix with the loss functions is filled (handwritten functions are in the other R-studio file) and the Diebold Mariano tests are done. After this for loop, all the values are calculated by taking the average of the accumulated values. The third file is used for the FTSE100 data. First, three different splits were used for the 50/25/25, 60/20/20, and 70/15/15 distributions. From there it is equivalent to the Monte Carlo simulation except for the for loop as it is only one run. Again, in the end, the metrics and statistics of the loss functions and tests are computed.