ERASMUS SCHOOL OF ECONOMICS

Finding effective weights to combine forecasts

A search for effective weights to combine volatility forecasts

Niels Holtrop

362040

Supervisor: Prof. dr. D.J. Van Dijk

Bachelor Scriptie Econometrie en Operationele Research

30 juni 2014

Abstract

We investigate the effectiveness of different weighting schemes to combine a set of forecasts from linear regression models. We use a set of 2^{18}=162.144 one-step ahead forecasts from a previous paper by Holtrop et al.(2014), as well as a selection of the best models, to all of which we assign a weight and create a single forecast. Our goal is to determine the effectiveness of different weighting schemes in comparison to the simple average of all the forecasts. We will evaluate the following schemes: The median, inverse MSPE weights, Bayesian Averaging, Principal Component Analysis and K-mean-clustering. The forecasts from all models will be evaluated statistically as well as economically, by creating a fictive investment strategy based on the produced forecasts. Both give somewhat different results, but the main findings are that the mean is a very solid benchmark and that other weights are most effective when using the full sample of forecasts. The inverse MSPE based weights perform the best of all the created weighting schemes for the full sample, and the K-Mean-clustering algorithm also gives promising results, especially because only a basic version of the algorithm was used. This can be an interesting weighting scheme for future research.

Keywords: Forecast combinations, volatility, inverse MSPE weights, Bayesian Averaging, Principal Component Analysis, K-Mean-Clustering

Contents

1. Introduction

The prediction of variables is a hot topic in today's economic circles; it has become more accessible to all who need it, and the research on finding the best model for each variable has become more widespread than ever. It is an essential subject for most companies, be it in terms of risk management, marketing purposes or logistical planning. From all ranges in the world economy, expectations of the behavior of the economic variables are needed to predict future market states and to determine new market strategies. A vast amount of literature has been dedicated to finding the best way to forecast each type of variable. Most of the research focuses on finding the best individual model, which can at best be only a reasonable approximation of the reality. Many methods require deep knowledge of the underlying structure of the variable that is being forecasted, most of which are never completely clear. Furthermore, many assumptions still need to be made and a robust model would be very useful.

For a simple forecast of a variable there are many different ways to approximate this value. Where one forecaster can assume a linear relationship, another one might prefer a non-linear model specification. The question arises which type of forecast is the correct one, and up to what degree one wants to spend his time to tightly specify the exact model. When multiple differently created forecasts are available, the ultimate goal is to extract as much information as possible from these forecasts. A search can be started to identify the best model, but Bates and Granger(1969) have suggested combining the forecasts obtained from the different models. One can argue that the combination of forecasts can offer diversification gains that make this choice very attractive. Even if the best model at each point in time could be identified, a combination of forecasts could still be attractive, even though its success would largely depend on the quality of the weights that are used to combine the forecasts (Timmermann, 2006).

Forecast combinations have proven to produce very good forecasts compared to advanced case-specific model specifications. Bates and Granger provide the following reasons for this; First, because the true data generating process is often unknown, even the best model is likely to be misspecified and often provides only a reasonable "local" approximation. Second, because the best model is likely to be time variant, a combination of models can provide a better forecast than a single model because there is no single model that is best over the entire horizon. By combining the models, the appropriate information of the best model can be incorporated at each moment in time. Last, forecast combinations offer diversification gains over regular models. Even if the best model could be identified at a point in time, a combination of models could still lead to a lower mean squared prediction error than that of the best specified model. The efficiency of a forecast combination can be motivated by the same idea as the creation of a portfolio of stocks, resulting in a lower overall risk than the risk of the safest stock.

The aim of this research is to find the best way of combining a set of forecasts. In a previous research done by Holtrop et al.(2014), a range of different model types was investigated with regard to their forecast quality for the realized volatility of a set of four asset classes, respectively stocks, commodities, bonds and an aggregated foreign exchanges measure. The research included a principal component analysis, as well as partial least squares regressions and forecast combination schemes with mean and median weights. The forecast combination method proved to provide very good results, often the best in terms of forecast quality. In this research, we will try to improve the

forecast combination method by looking at more advanced weighting schemes, where we use the mean forecast as a benchmark to compare against.

The forecasts for the forecast combination scheme were constructed by creating all possible linear models with 18 macroeconomic- and financial variables and an autoregressive term. This resulted in 2^{18} = 262.144 models, which were then used to produce one-step-ahead forecasts. In this research, all the results with respect to the first asset class, stocks, will be presented. Results for the other 3 asset classes showed similar results, and are available on request. For a uniform evaluation, the forecast are evaluated from forecast 26 and onward, because some models start forecasting from this period onward.

The weighting schemes that will be used will both be performed on the full set of 2^{18} forecasts, as well as on a selection of the top models. Previous research has shown that the simple average of produced forecasts often provides a very good forecast relative to more advanced weighting schemes. Prior papers however often analyzed a smaller set of forecasts, with forecasts from different types of models, resulting in a small amount of forecasts relative to that of this research. In that case, the forecasts are often more uniformly spread around the true value than is the case with solely these linear models in this paper, due to which more advanced schemes often proved redundant. Due to the large quality difference in the forecasts used in this paper, schemes based on forecast-model quality are likely to perform better than the mean, because selecting the good forecasts can bring more gains than with a small set of forecasts as used in previous research.

The mean forecast will be used as the benchmark to be beaten in this research. Weights based on model quality will be both based on in- and out-of-sample performance. The first scheme that will be used is the median of all forecasts, because the median is more robust to outliers. The second scheme is based on each model's out-of-sample performance, as measured by the inverse mean squared prediction error (MSPE) relative to that of the others. In the previous research of Holtrop(2014) it was found that the performance of the forecasts varies quite substantially over different periods, so a weight based on the moving MSPE could provide better fits over all periods. The third model builds on in-sample performance, in the sense that it determines which model is most efficiently specified based on the Bayesian Information Criterion (BIC), which uses the in-sample variance as an estimator for the models' variance. A small in-sample variance follows from a smaller spreads between the true- and estimated value in the estimation period, which in turn implies a better forecast quality because of less uncertainty regarding its value. The fourth scheme uses a factor based approach to create linear combinations of the forecasts, which are then used in a linear regression, reducing the complexity of the problem. The fifth and final scheme is the K-Mean algorithm, which splits the forecasts into K clusters based on their forecasting performance. It could be that the first cluster of forecasts encompasses the information of all the forecasts from the other K-1 clusters, due to which the first cluster will be sufficient for forecasting.

We have found that in sets where the amount of forecasts is very big, there are gains to be found when using more advanced weighting schemes. The inverse MSPE based weights, as well as the Bayesian Averaging scheme and the K-Mean-Clustering give better results than the mean. The K-Mean algorithm that has been used has a very basic specification and can possibly be advanced further to provide even better results in future research.

This research is organized as follows. Chapter 2 describes the data which is used. In Chapter 3 the methodology used is discussed. Chapter 4 contains the results which are evaluated both statistically and economically. Finally, in Chapter 5 the conclusion is presented.

2. Data

In this research the results from previous research by Holtrop(2014) will be used, which in turn are based on a dataset with *T=336* monthly observations of 38 macro economical and financial variables by Christiansen et al.(2012) from January 1983 up to December 2010. The latter dataset also contained the volatility of exchange rates, commodities, stocks and bonds for each month, which have been used as dependant variables in the predictive regressions. These volatilities are defined as the natural logarithm of the square root of the realized volatility, which in turn is estimated as the sum of the squared daily returns. A description of the basic statistical properties of all the explanatory variables as well as of the volatility is attached in Appendix A. In this report, all the results will be presented for the stocks; the results for the other three classes are available upon request. For a more in-depth analysis of the variables, their effect on the volatility and the construction of the forecasts we would like to point to the paper of Holtrop (2014).

The regressors for the regressions have been extracted from the bigger set of 38 variables by means of the Least Angle Regressions (LARS) algorithm. This LARS algorithm ranks the variables by means of their correlation with the dependant variable and thus results in the relevance of the variables in explaining the volatility. Because the model specification with the top 18 of all 38 variables and a moving window of 10 years proved to be the best specification in the research of Holtrop(2014), we will only forecasts produced by this model specification in our paper. With these 18 variables, all possible linear regression models have been created, to all of which an autoregressive term of the volatility was also added since this carries a lot of information, resulting in a total of 2^{18} = 262.144 models. Next one-step-ahead forecasts were made with each of these models, resulting in 2^{18} one-month-ahead forecasts of the volatility for each of the 216 forecast periods.

The forecasts were created using a moving window of length T_1 = 120 months, the first estimation period being January 1983 up to December 1992 to forecast the first value in T_1+1 = January 1993, and the last period being December 2000 up to November 2010 to forecast the 216^{th} value in December 2010.

3. Methodology

The different weighting schemes will be performed on the matrix with all one-step-ahead forecast from all the created models, which is a set of 2^{18} = 262.144 forecasts for each of the 216 forecast periods. Later on, the forecasts will also be split up into a selection of the top 25 and top 100 forecasts to check the influence of this on the forecast quality and the performance of the different weighting schemes.

In Timmermann(2005), steps are shown to find the optimal weights for a set of forecasts. A derivation of the optimal weights for a combination of forecasts is given in Appendix C1. For the weighting schemes of which the weights are known, a comparison will be made with respect to the optimal weights by inserting both the weights in the formula for the model variance, and then dividing the two variances to check the relative performance of the weights. In what follows we will give a specification of the weighting models used in this paper and provide a motivation for their use.

3.1 Simple average

The benchmark weighting scheme results from taking the mean of all the N one-step-ahead forecasts made from all the models *j*.

$$\hat{y}_{t+1,t} = \frac{1}{N}\sum_{j=1}^{N} \hat{y}_{t+1,t,j} \ , \qquad t = T_1+1,\ldots, T$$

The simplicity of its specification makes it a very useful scheme to use in practice. The motivation for the use of the mean is very straightforward; suppose on has two forecasts on hand, both made by experts whom we don't know, with no further information on the relative quality of both forecasts. A good way to incorporate the information from both forecasts would then be to average them. A same reasoning supports the use of the simple average in the context of combining many forecasts from all possible linear models, where we don't know the quality of the individual forecasts made by each regression model.

In practice, the simple average has proven to be a very solid benchmark, which has been shown to be very hard to beat by more advanced schemes. Palm and Zellner(1992) gives three main advantages of using a simple average forecast;

1. "Its weights are known and do not have to be estimated, an important advantage if there is little evidence on the performance of individual forecasts or if the parameters of the model generating the forecasts are time-varying.
2. In many situations a simple average of forecasts will achieve a substantial reduction in variance and bias through averaging out individual bias.
3. It will often dominate, in terms of MSE, forecasts based on optimal weighting if proper account is taken of the effect of sampling errors and model uncertainty on the estimates of the weights."

Timmermann(2006) notes that the performance of the equal-weighted forecast combinations critically depends on the fact that all forecast errors need be of the same magnitude, their ratio being close to unity, and the correlation of forecast errors between model-pairs should also be roughly equal. Gupta and Wilton(1987) have found that different weighting schemes perform better when there are large differences between the sizes of the forecast errors from the different models.

In this research, all possible linear models are used to forecast the volatility. The assumption of the forecast errors for all models being of equal size is not very realistic. In Appendix B, the distribution of the forecast errors of all 2^{18} models for a few periods are shown, and we see that the individual forecasts aren't uniformly spread around the true value, from which we conclude that the models have different variances. Division of the variance of the mean scheme model on that of the optimal weights results in[1]

$$\frac{\sigma^2_{mean}}{\sigma^2_{opt}} = \frac{(\sigma_1^2 + \sigma_2^2)^2 - 4{\sigma_{12}}^2}{4\sigma_1^2\,\sigma_2^2(1-\rho_{12}^2)}$$

This is larger than one unless $\sigma_1 = \sigma_2$, from which we can conclude that the mean is only optimal given that the variances of the different forecast models are equal to each other. Having seen the plot in Appendix B, indicating that this is not the case, we can conclude that the possibilities for the more difficult models are open.

3.2 Median

The first weighting scheme takes the median of all the N forecasts:

$$\hat{y}_{t+1,t} = \text{median}_N\left(\hat{y}_{t+1,t,N}\right), \qquad N = 1,\ldots, 2^p \;,\; t = T_1{+}1,\ldots, T$$

With 262.144 models for each time period, of which we have no clue about relative predictive quality, many things are possible. Since all possible models are used, including one with only a constant and the autoregressive term, some of the forecasts can give results that lie far from the true value. Be it positive or negative outliers, the simple average can be negatively influenced by this. If for instance there are many forecasts that estimate (far) below the true value, the simple average will also be lower than the true value, resulting in a worse performance than one would want. The median is more robust to these outliers, and could thus give a better forecast.

If the forecasts errors are standard normally distributed, which means that they are spread uniformly over a certain interval around the true value, both schemes will result in a roughly equal value. When this is no longer true, Marcellino(2004) has shown that linear combinations of the forecasts are no longer necessarily optimal. For a very much fluctuating series as we use in this paper, it is useful to look at both schemes. Computational simplicity of this scheme also supports its use in simple forecasting applications.

3.3 MSPE based weights

The second weighting scheme is more advanced, in the sense that it involves the individual predictive quality of each model. It has often times been argued that it is very difficult to precisely estimate the covariance matrix of the forecasts errors, because the amount of forecasts to combine is often substantially large. Stock and Watson(2001) proposed to ignore the correlation between the models and base the weights only on the models' relative performance, as measured by their Mean Squared Prediction Error (MSPE). To get the most out of this scheme, it is more relevant whether a model performed well over recent periods than how it has done over the entire past period. To accomplish this, the MSPE will be calculated with a moving window of length 24 months, resulting in weights that are based on current performance.

[1] For derivation, see Appendix C.2

The mean squared prediction error with a moving window of 24 months for model *i* at time t is defined as

$$MSPE_{i,t} = \frac{1}{24} \sum_{j=t-24}^{t-1} (y_j - \hat{y}_{j,i})^2 \text{ ,} \qquad i = 1,...,N \text{ , } t = T_1+26,....,T$$

The MSPE for model *i* at time *t*, where T_1+1 refers to the first made forecast at 1993.1, is calculated as the average squared forecast error over the past *24* forecasts, resulting in a moving window of length 24 months. This means that the first weight that can be estimated is that of the 26th forecast, because we only have MSPE values for period 2 and onward. For instance at time $t=T_1+26$, the weights are based on the MSPE that is taken as the average of the squared prediction errors of forecasts 2 to 25. The weights for each forecast are then defined as the inverse mean squared prediction error raised to a power k divided by the sum of all inverse MSPE's raised to the same power *k*.

$$w_{i,t} = \frac{MSPE_{i,t}{}^{-k}}{\sum_{j=1}^{N} (MSPE_j, t^{-k})} \text{ ,} \qquad i = 1,...,N \text{ , } t = T_1+26,....,T$$

This scheme will result in larger relative weights for better performing models since these have a lower MSPE, which in turns becomes large as we take inverse powers. This parameter *k* changes the relative size of the weights. Since MSPE values are often in the same order of magnitude, with a maximum difference of around a factor 10, the regular inverse MSPE(*k=1*) gives weights that are still all close to each other. To create a bigger spread between the weights, where better forecasts get a significantly bigger weight, choices of *k>1* can be used. In this paper, we will partially follow the paper by Marcellino(2004) and look at the weights with k=1, 2, 5.

The MSPE based weights work well in practice because the off-diagonal elements of the covariance matrix of the forecast errors don't have to be estimated. An important paper by Timmermann (2005) on the diversification gains of different weighting schemes under quadratic loss shows that these weights are only optimal in large samples provided that the correlation between the forecast errors is truly equal to zero. Given that we have two sets of forecasts with corresponding model variances ${\sigma_1}^2$ and ${\sigma_2}^2$, the ratio of variance of the MSPE weights and the optimal weights under quadratic is given by[2]

$$\frac{\sigma^2_{MSPE}}{\sigma^2_{opt}} = \left(\frac{1}{1-\rho^2_{12}} \right) \left(1 - \left(\frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2} \right)^2 \right)$$

which is larger than one unless $\rho_{12} = 0$ when $\sigma_1 \neq \sigma_2$. We can thus conclude that the MSPE based weighting scheme is only optimal given that there is no correlation between the used forecasts.

[2] For derivation, Appendix C.3

3.4 Bayesian weights

Buckland et al.(1997) propose to base weights on a Bayesian procedure. Ex ante any of the *N* models is equally likely to be the best model for predicting y_t, and they show that

$$w_{i,t} = \frac{\exp\left(-\frac{1}{2}\Delta BIC_{\mathrm{i,t}}\right)}{\sum_{j=1}^{N}\exp\left(-\frac{1}{2}\Delta BIC_{j,t}\right)}\ , \qquad i = 1,\ldots,N\ ,\ \ \mathrm{t} = T_1+1,\ldots,T$$

gives the posterior odds of model *i* being the best predictive model. Here *ΔBIC* is defined as the value of the Bayesian Information Criterion of model *i* minus the minimum of all the BIC values, with BIC_i defined as

$$BIC_{i,t} = T_1 \ln \hat{\sigma}^2{}_{i,t} + (m_i + 1)\ln T_1\ , \qquad i = 1,\ldots,N\ ,\ \ \mathrm{t} = T_1+1,\ldots,T$$

Here, m_i denotes the number of variables used in the *i*-th regression model, ranging from 1 for only the autoregressive term to 19 for all included variables, and T_1 is the number of observations in the estimation sample, which equals 120 for all models due to the moving window model specification. Adding more variables to a regression increases the estimation error, but could increase the forecast quality. This criterion weighs the addition of extra predictive variables against the effect on the models variance σ^2, which is estimated by the in-sample variance which is calculated as

$$\hat{\sigma}^2{}_{i,t} = \frac{1}{T_1-1}\sum_{j=t-T_1}^{t-1}\left(y_j - \hat{y}_{i,j}\right)^2, \qquad i = 1,\ldots,N\ ,\ \ \mathrm{t} = T_1+1,\ldots,T$$

Here, y_j is the observed true value at time *j* and $\hat{y}_{i,j}$ is the predicted value of model *i* for time *j* according to the least squares regression line. Values of the BIC are always negative, and a smaller negative value (i.e. closer to zero) implies a better model quality. The difference between BIC_i and the best model, *ΔBIC*, will thus be a positive value. The weights are then defined as the exponent of minus a half times the positive value, resulting in the exponent of a large negative value for forecasts made by relatively bad models, and the exponent of zero for the best model. The interval of the different weights values is thus very different from the previous weighting schemes, because good models get very large weights compared to the worst models.

Diebold(1991) notes that the performance of the Bayesian Averaging method mainly depends on the validity of a maintained assumption, namely that one of the models whose forecasts are combined is the true data generating process (DGP). When the true data generating process is among the models to combine, this model gets a very large relative weight, with the numerator being equal to $e^0 = 1$ versus e to a (large) negative power for all other forecasts, which is very beneficial for the schemes forecast quality. However, when the true DGP is not among the models, the best specified model will still get a weight of one in the limit, but the parameter estimation will converge to a pseudo-true value, a value which is closest to that of the true DGP. Since this model is false however, we will not make a good parameter estimation. The problem that arises here is that the true reason of forecast combination methods, which is to obtain diversification gains from combining information from multiple forecasts, will no longer be applicable in this case because most of the information in the other forecasts is being discarded.

3.5 Principal Component Analysis

To reduce the dimensionality of the problem, a factor based approach seems suitable. Hsiao and Wan(2014) propose to use an eigenvector based approach, performed on the matrix containing all forecast errors for a single time period. Next, weights would be determined by finding the minimum value over the set of each eigenvalue divided by the sum of the elements of its respective eigenvector, and setting the weight equal to the eigenvector following from this procedure, and then dividing each element by the sum of all the elements. Because of the vast amount of forecasts, the eigenvector based approach as proposed by Hsiao and Wan isn't feasible in this context. Often times there will only be a single unique eigenvalue, due to which we can't perform the algorithm proposed in their paper.

Another factor based approach is Principal Component Analysis (PCA), which should be able to incorporate the information of all forecasts in an efficient way. This technique involves describing the structure of the variation in the dataset in terms of a set of uncorrelated factors, with every factor being a specific linear combination of the original forecasts[3]. The linear combinations that turn out to maximize the variance are the eigenvectors from the correlation matrix, ordered from the one belonging to the largest eigenvalue up to the one with the smallest eigenvalue. In this dataset the eigenvalues will almost all be zero due to the size of the correlation matrix, which will be far from full rank. The few eigenvalues that are not equal to zero will however be enough to explain enough of the variance in the set of forecasts.

These factors will then be regressed on the real values of the volatility by means of OLS, and forecasts can then be made. In this model again, the first 24 observation are used to initialize the model, so forecasting starts from forecast 25 and onward. Factors will be added until a minimum of 80% of the total variance is explained, which will often already be accomplished by only the first factor, because of the structure of this dataset. We will apply this analysis to a subset of all models, because it isn't feasible to create the sample correlation matrix for all the 262.144 models. This subset will consist of the maximum computationally feasible number of forecasts, which equals 3000 for our setup.

3.6 K-mean-clustering

The final scheme that will be used is the K-Mean-Clustering algorithm, proposed by Aiolfi and Timmermann(2006). In their paper, the forecasts are split up into K clusters based on their historical MSPE performance, ranging from cluster 1 with the lowest MSPE's up to cluster K with the highest MSPE's. For all these K clusters, the mean is taken as a regressor, resulting in K regressors to be regressed on the true value y_{t+1}. The choice for K is however of great importance, because the splitting should be based on some objectively defined property. In this paper, we will choose K based on the mean silhouette value, inspired by a paper on K-mean-clustering by Baridam(2012).

In this scheme, the observations of the sample are split into *K* clusters by means of a distance objective function that has to be minimized. Observations are plotted with the squared error at time t+4 against the mean squared forecast error from *t* to *t+4*[4], and then divided into clusters according to their similarity, with the interpretation of respective forecast quality per cluster being very straightforward. The group similarity is measured by the distance to its respective cluster centroid,

[3] For an in-depth analysis of PCA, we refer to the paper by Holtrop et al.(2014)
[4] See Appendix E for example

and different choices can be made as distance measures, ranging from squared Euclidian distance and the cosine measure, to a sample correlation related measure. We will go with the most common one, being the squared Euclidian distance. The performance of the split can be analyzed by means of the silhouette value q_i. The silhouette value is a measure of the fit to a cluster, weighing within-group fit against fit with the other clusters.

$$q_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, -1 < q_i < 1$$

Here, a(i) is the average similarity between x(i) and the rest of the points assigned to cluster K^j, as measured by the distance to the centroid of K^j , and b(i) is the minimum average similarity between x(i) and the rest of the object in all the other $K \neq K^j$ clusters. If we calculate this for every point *i* in every created cluster K^j and take the mean value of all these points, we can see if the clustering was useful. A value of q_i equal to one implies a perfect fit, a value of zero implies that a point can belong to any of the clusters and a value of minus one indicates that it is in the wrong cluster..

For a set of values for K ranging from 2 to 7, the dataset will be divided into K clusters. Each data point is then grouped into a cluster *k*, and all points are then compared to their respective group members and the points in the other *k-1* clusters. This results in silhouette values for all points, and the mean silhouette point is then taken as a statistic, and plotted against the value of K, creating an elbow plot like figure, where to value of K is chosen which produces the highest mean silhouette value.

Next, the mean of all forecasts in each cluster *k* will be used as a regressor in a least squares regression. The value of K is thus best chosen not to large, because for large K we would increase the estimation error in the regression. We will create models with and without the intercept, and compare them since the coefficient could compensate for a present bias in the forecasts.

$$y_{t+1} = \alpha + \sum_{i=1}^{K} \beta_i \mu_i + \varepsilon_{t+1}, \qquad t = T_1+25,\ldots,T$$

We initialize the model with the first 24 observations, and then start estimating the regression coefficients with a moving window of length 24. The forecast is then made by multiplying the coefficients β_i with the observed values of the respective regressors. This model starts from forecast 25 instead of the regular first (mean & median), as was also the case for the MSPE weighting scheme. Because of the computational complexity of current silhouette value approximations, the analysis is performed on the top 1000 models based on their average MSPE over all 216 forecasts, instead of on the full sample of forecasts.

Because of the big difference in performance between the forecast models, splitting the forecasts based on their performance could lead to an increased forecast quality. It could also be interesting to see if only the best few clusters get weights, and the worse clusters simple get a coefficient of zero. This would imply that the good forecasts encompass all the information from the lesser forecasts.

3.7 Statistical Analysis

To compare the relative performance of the models, a statistical analysis will be performed. For a uniform comparison, we will start the statistical analysis at forecast 26 instead of forecast 1 since some models use a moving window for the estimation, where the first set of observation is used as an initialization period.

First we will check whether the models produce unbiased forecasts, which is one of the main properties that a good estimator should possess. This property means that the expected value of the forecasts that are produced by the model should equal to the true value of the variable that is being estimated. This hypothesis can be tested by a standard Z test, defined as

$$Z = \frac{X - \mu}{\sigma}$$

where X is the estimated value, μ is the true value and σ^2 is the variance of X. This statistic follows a standard normal distribution, which results in a critical value of ±1.96 at a 5 percent significance level.

The second value that will be reported is the Mean Squared Prediction error (MSPE) of each model, relative to that of the mean weight scheme. This gives an indication of the relative performance of a model, with an MSPE smaller than one for a better forecast which is better than that of the mean scheme. Since these aren't directly comparable in the sense that a lower MSPE doesn't automatically imply a better forecast, the Diebold Mariano(1995) test will be used to compare different forecasts, which will be explained later in this section.

To check for the efficiency of the forecast, Mincer and Zarnowitz(1969) proposed to check this by performing a regression of the forecasted value and a constant on the realized value.

$$Y_{t+1} = \alpha + \beta \hat{Y}_{t+1} + \varepsilon_{t+1}$$

A Wald test is then performed on the joint restriction of both coefficients, with the restrictions α=0 and β=1, the null hypothesis being optimality of the forecasts under MSPE loss. The idea behind these restrictions is that there should be no way of being able to predict the forecast error at a time *t+1* when knowing $\hat{Y}_{t+1}$. The level of significance will again be 5 percent.

The relative forecast quality of the different models will be assessed by means of the Diebold-Mariano test. Say we have two sets with each *P* one-step-ahead forecasts for our series, say from model *i* and model *j*, for times *t=T,...,T+P-1.* Define the "loss differential" as $d_{t+1} = {e_{t+1|t,i}}^2 - {e_{t+1|t,j}}^2$. Equal forecast accuracy implies that the expected value of the loss differential is equal to zero. We can test for this by means of the Diebold-Mariano test statistic, given by

$$DM = \frac{\bar{d}}{VAR(\frac{d_{t+1}}{P})}$$

which is distributed standard normal. Here $\bar{d}$ is the sample mean of d_{t+1} and Var(d_{t+1}) can be estimated by $\frac{1}{P-1}\sum_{T}^{T+P-1}(d_{t+1} - \bar{d})^2$. If the absolute DM test statistic is larger than the critical value, we reject the null hypothesis of equal forecast quality. The one sided alternative hypothesis leads to comparative conclusions; if the DM statistic is positive, this means that $\bar{d}$ is positive, which in turn means that ${e_i}^2$ is larger than ${e_j}^2$, and thus that model *i* produces worse forecasts because it has larger forecast errors.

3.8 Economical Evaluation

A new type and practical type of evaluation which is more and more used is the economical evaluation. For this analysis, we will follow Cakmakli and Van Dijk(2013). Here, a fictive investor will base his portfolio selection on an investment function. The investment function of returns and volatility forecasts is given by

$$\max_{w_{t+1}} \; E(r_{p,t+1}) - \frac{1}{2}\gamma Var(r_{p,t+1})$$

Here γ is the rate of risk aversion. The portfolio return ($r_{p,t+1}$) consists of a risk-free 3-months T-bond return and monthly stock return of the S&P 500($r_{s,t+1}$). For the economical evaluation of the forecasts for the volatility of the stocks, the portfolio consists of the risk free return and stock returns. Unfortunately no data is available for 1-month T-bill rate before year 2001, so therefore the 3-months t-bill rate is used for the risk-free return ($r_{f,t+1}$). The portfolio return ($r_{p,t+1}$) is given by

$$r_{p,t+1} = (1 - w_{t+1})r_{f,t+1} + w_{t+1}r_{s,t+1}$$

The return forecast will be the average return of the past 10 years, the variance forecast will be a transformation of the forecasts that are constructed in this paper. To obtain the optimal weights in the portfolio, the investment function is maximized. The variance of the risk free return is equal to zero because it is assumed that $r_{f,t+1}$ is fixed at the end of month t. For the variance of the portfolio it holds that

$$Var(r_{p,t+1}) = w_{t+1}^2 Var(r_{s,t+1})$$

So, the optimal weights w_{t+1}^* can be derived by maximizing equation (6), which are given by

$$w_{t+1}^* = \frac{E(r_{s,t+1}) - r_{f,t+1}}{\gamma Var(r_{s,t+1})}$$

Two cases are considered; first, the weights are bounded between zero and one ($w_{t+1}^* \in [0,1]$). These weights imply that short selling and lending are not allowed. In the second case, short selling and lending are permitted ($w_{t+1}^* \in [-1,2]$). Transaction costs are neglected. To evaluate what an investor is willing to pay for using the volatility forecasts of this paper, the maximum performance fee is calculated. To be able to do this, a quadratic utility function is assumed (West, Edison & Cho, 1993). The average utility is given by

$$\bar{U} = \frac{W}{n}\sum_{t=0}^{n-1}\left(R_{p,t+1} - \frac{1}{2}\frac{\gamma}{(1+\gamma)}R_{p,t+1}^2\right) \qquad with\ R_{p,t+1} = 1 + r_{p,t+1}$$

Here W is defined as the wealth to be invested and *n* is the number of time periods where the investing is analyzed. In order to calculate the maximum performance fee, the utility of a strategy arising from the forecast of the constructed models (*strategy a*) needs to be compared with an unsophisticated buy-and-hold strategy (*strategy b*). The buy-and-hold strategy consists of either only investing in the risk-free t-bonds, only investing in the market, or in an equally weighted combination of the two.

$$\sum_{t=0}^{n-1}\left((R_{p,t+1}^a - \Delta) - \frac{1}{2}\frac{\gamma}{(1+\gamma)}(R_{p,t+1}^a - \Delta)^2\right) = \sum_{t=0}^{n-1}\left(R_{p,t+1}^b - \frac{1}{2}\frac{\gamma}{(1+\gamma)}(R_{p,t+1}^b)^2\right)$$

From this equation the delta can be calculated, which is a fraction of the wealth that the investor is maximally willing to pay for this information. The higher this delta, the better a model performs, since the investor makes a profit of at least (delta*100) percent.

4. Results

To analyze the performance of the different weighting schemes from Chapter 3, we will analyze the results for all the models. In Section 4.1, the basic statistical properties will be highlighted to indicate their individual performance. In Section 4.2 the models will be compared with their counterparts; Section 4.2.1 compares each model to the relevant benchmark model, Section 4.2.2 compares the different models with each other, which will lead to a ranking of all the models. Section 4.3 encompasses the economical evaluation, for which the realized variance will be used. The transformation from realized volatility to realized variance is included in Appendix H.

4.1 Statistical properties

We will first analyze the quality of the different weighting schemes via a statistical evaluation, for which the results are shown in table 1. The models are tested for at unbiasedness at a level of 5%, resulting in a critical value of 1.96. The MSPE values are displayed relative to that of the mean scheme which corresponds to the same set of forecasts for a fair comparison.

Model	Unbiasedness	MSPE	Mincer-Zarnowitz
Mean	-2,64*	1,000	0,0311*
Mean, top 25	-2,46*	1,000	0,0227*
Mean, top 100	-2,35*	1,000	0,0361*
Median	-2,67*	1,088	0,0295*
Median, top 25	-2,51*	1,113	0,0146*
Median, top 100	-2,37*	1,123	0,0306*
MSPE expanding window	-2,65*	3,314	0,0304*
MSPE moving window, k=1	-2,66*	0,995	0,0305*
MSPE moving window, k=2	-2,68*	0,990	0,0296*
MSPE moving window, k=5	-2,73*	0,983	0,0261*
MSPE moving window, top 25, k=1	-2,45*	0,990	0,0230*
MSPE moving window, top 25, k=2	-2,44*	0,992	0,0225*
MSPE moving window, top 25, k=5	-2,40*	0,998	0,0210*
MSPE moving window, top 100, k=1	-2,35*	0,998	0,0283*
MSPE moving window, top 100, k=2	-2,36*	1,000	0,0332*
MSPE moving window, top 100, k=5	-2,35*	0,998	0,0283*
Bayesian Averaging	-2,13*	0,998	0,0304*
Bayesian Averaging, top 25	-2,05*	1,075	0,0860
Bayesian Averaging, top 100	-2,84*	1,107	0,0037*
PCA, top 3000	0,14	0,098	0,0152*
PCA, top 25	0,19	0,907	0,0649
PCA, top 100	0,20	0,907	0,0528
K-mean with intercept	0,59	1,041	0,0284*
K-mean without intercept	-0,31	1,016	0,0125*
K-mean with intercept, top 100	-0,05	1,098	0,0226*
K-mean without intercept, top 100	-0,73	1,078	0,0220*

Table 1. Information based on basic statistical tests.

Note: Displayed are the results for the mean, median, inverse MSPE based weights values of k being 1,2 and 5, the Bayesian Averaging model, Principal Component Analysis and the K-Mean-Clustering algorithm, all applied to the full sample as well as the selection of top 25 and top 100 forecasts. The second column contains Z statistics for unbiasedness. The third column contains the relative MSPE values, e.g. for PCA top 25 being the ratio of the MSPE of PCA top 25 divided by that of mean top 25. The fourth column contains the p value for the Mincer-Zarnowitz test. An asterisk indicates significance at 5 percent level.

Almost all models give negatively biased forecasts, indicating that they all tend to overestimate the true value in a negative way. The only models that produce unbiased forecasts are the Principal Component Analysis and the K-mean algorithm. These two schemes seem to weigh the forecasts very well, and are capable to remove any bias present.

The Mean Squared Prediction errors (MSPE) of the models are all roughly equal, the only clear outlier being the MSPE of the MSPE expanding scheme. As was said before, the performance of the expanding MSPE scheme is most likely due to its bad adjustment to relative performance of models, resulting in forecasts that aren't time optimal. All other models seem to forecast very well, with a good average quality of their forecasts. The results for unbiasedness make clear that the models often seem to forecast below the true value, but we can see that the forecasts lie close to the real value.

In column four, the p values for the Mincer-Zarnowitz regressions are displayed. For almost all models, we reject the null hypothesis of forecast optimality under MSPE loss at the 5 percent level, so we can conclude that all models produce inefficient forecasts. The only model that produces efficient forecast is the PCA model on the selection of top 25 and top 100 models and the Bayesian averaging model on the top 25 selection. This result seemed to be due to the length of the evaluation period, since for evaluation from forecast 1 until 216 the forecasts for the relevant schemes were efficient and unbiased. An explanation for this is still missing.

4.2 Model Comparison

For each model, different specifications have been used during the estimation. The question is up to what degree this increases the performance of the weighting scheme. Regressions were first run on all the available 262.144 forecasts, which contained many forecasts that were off. In Appendix D the spread of all the forecasts for one time period is shown, and it is clearly visible that the spread among the forecasted values can be humongous. With the true value at around -1.43, model estimates range from zero to -2,5. A preselection was made of the forecasts, resulting in a top 25 and top 100 of models. This greatly increased the average quality of the forecasts and drastically decreased the amount of forecasts. The spread between these forecast is now very small, as can be seen in figure 2 and 3 in Appendix D, resulting in a selection from on average better forecasts. In the next two sections, the differences between the specifications will both be analyzed for individual weighting schemes as well as between all the schemes.

4.2.1 Intra-Model Comparison

We start off with results for the basic mean and median based weighting schemes. The Diebold-Mariano statistics for these forecasts are shown in table 2 and 3. For both models, the preselection of models results in a better quality of the forecasts. As for both the mean and the median the quality of the top 100 models is significantly better than that of the top 25, we can conclude that the performance increases when the amount of models included in the top models increases. This is quite unexpected, since one would expect that the simple average of the top 25 models would perform better

Model	Mean	Mean top 25
Mean	-	-
Mean top 25	-1,757*	-
Mean top 100	-2,164*	-1,711*

Table 2. Diebold Mariano statistics for comparison of relative forecast quality for the Mean scheme. An asterisk indicates significance at 5 percent level.

Model	Median	Median top 25
Median	-	-
Median top 25	-1,295	-
Median top 100	-1,807*	-2,390*

Table 3. Diebold Mariano statistics for comparison of relative forecast quality for the Median scheme. An asterisk indicates significance at 5 percent level.

than that of the top 100 models. An explanation for this result is still missing, but could have something to do with the negative biasedness of all forecasts, where the top 25 are all biased one way, whereas the top 100 are more spread, some being biased upward and some downward. For all models, the mean significantly outperforms the median weights, so we use the mean as the benchmark for the other models. The outperformance of the top 25 by the top 100 gives rise to the question whether it would be beneficial to include more models in the top selection. In appendix F, we have included a set of DM-statistics, which show that the outperformance stops at the top 100, and this is thus the most beneficial selection.

In table 4 the results for the MSPE based weights are shown in comparison to those of the mean. In table 4, the first thing that we note is that the MSPE weights with the expanding window perform very poor, as was expected. Furthermore we see that the regular inverse MSPE scheme outperforms the mean weights, but for the other cases of k this is not the case. For the selection of top 25 and 100 models there is no significant difference between the qualities of the forecasts. For this scheme, we conclude that the increase of the value of k does not provide us with a better forecast quality.

Model	**Mean**
MSPE expanding	2,021*
MSPE k=1	-1,663*
MSPE k=2	-1,508
MSPE k=5	-0,950
Model	**Mean top 25**
MSPE top 25 k=1	-1,222
MSPE top 25 k=2	-0,903
MSPE top 25 k=5	-0,152
Model	**Mean top 100**
MSPE top 100 k=1	-1,222
MSPE top 100 k=2	-0,903
MSPE top 100 k=5	-0,152

Table 4. Diebold Mariano statistics for comparison of relative forecast quality between the inverse MSPE schemes and the respective mean weights. An asterisk denotes significance at 5 percent level.

In table 5, the results for the Bayesian Averaging scheme are displayed. The Bayesian Averaging model barely benefitted from the preselection. This can be partially explained by the main idea behind this scheme, which was that it gives a very big weight to the best specified model, and almost zero weight to the rest of the forecasts. In the full set of forecasts as well as in the selection, it will select the same model and assign the rest of the weight to the other forecasts, resulting in very little difference between forecasts from the two different selections. There was no significant difference between the forecast quality of the Bayesian averaging model and the mean when performed on the full sample. On the selection of top 25 and 100 forecasts, the Bayesian averaging model had significantly worse forecasts

Model	**Bayes**	**Bayes top 25**
Bayes	-	-
Bayes top 25	-0,453	-
Bayes top 100	-0,580	-1,007

Table 5. Diebold Mariano statistics for comparison of relative forecast quality for the Bayesian weight scheme. An asterisk indicates significance at 5 percent level.

For the Principal Component Analysis, the results are presented in table 6. The PCA weights are not able to provide better forecast than the mean for any of the selections. This is probably caused by the estimation of the correlation matrix of the forecasts, which will have large estimation errors due to its size, as well as the very solid performance of the mean. With forecasts all having good quality here, there is very little room for improvement upon the mean of all selected forecasts.

Model	**Mean**	**Mean top 25**	**Mean top 100**
PCA top 3000	-0,269		
PCA top 25		0,363	-
PCA top 100		-	0,688

Table 6. Diebold Mariano statistics for comparison of relative forecast quality for the PCA weight scheme. An asterisk indicates significant at 5 percent level

The k-mean algorithm was performed on the top 1000 models with the lowest average MSPE. For the choice of K, the plot of the mean silhouette values is displayed in figure 1. We can conclude that we should choose two clusters for this sample, because this is the only peak. Next, the mean value of each of the two clusters was used as a regressor on the true value, in a regression with and one without an intercept. In figure E.2 and E.3 in Appendix E, we have plotted the estimated regression coefficients over time, for both the regression with and without an intercept. We see no clear patterns in their course, and the coefficients vary quite substantially over time. We see that the intercept is often different from zero, indicating that a bias is presumably being adjusted. Also we see that the second (third) coefficient β_2 shows a same pattern as the first (second) coefficient β_1 and is not equal to zero, as would have been the case if the forecasts from the first cluster encompassed the information in the second cluster. The Diebold-Mariano statistic showed no significant differences between the quality of the forecasts from both models.

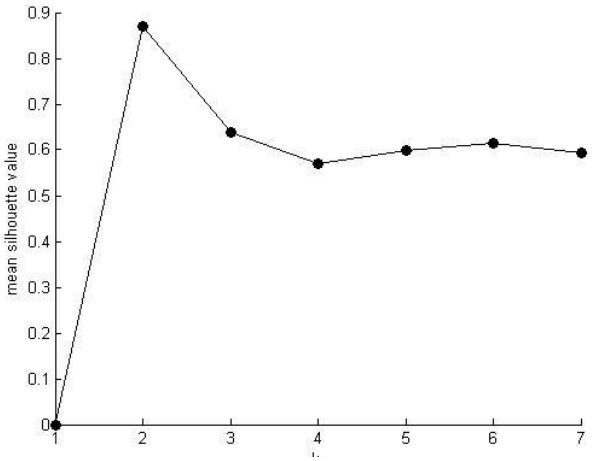


Figure 1. Graph of the value of K against the mean silhouette value. ***Note: A higher mean silhouette value implies a better overall cluster fit.***

4.2.2 Inter-Model Comparison

Having seen that the preselection offers significant gains for all weighting schemes individually, the question arises what the effects are on the relative quality change for the different schemes. We will first analyze the performance of the models over the full range of 262.144 forecasts, of which a plot for 1 time period was shown in figure D.1.

Because of the large spread of the forecasted values, a good selection method can offer many gains. The big relative differences between the individual forecasts give good options for the weighting schemes that choose the right forecasts, leading these to predict a more accurate value than that produced by taking a simple average. As was seen in figure D.1, the mean and median of all forecasts probably offer a forecast that is biased upwards. With a better weighting scheme, only the values that offer good forecasts will be given large weights, resulting in an increase in forecast quality.

In table 7 results are shown for the relative quality of all the models over this full sample of forecasts. It can be seen that the MSPE based weights produce significantly better forecasts than the simple average and the median of all forecasts. This complies with the expectation that was stated earlier, because the MSPE increases the weight of models that perform well. The performance of the Bayesian Averaging weighting scheme is somewhat disappointing. This mediocre performance can probably be explained by the violation of the assumption that is made in this

Model	Mean	MSPE	Bayesian	K-Mean
Mean	-	-	-	-
Median	0,512	1,781*	0,641	-1,652
MSPE k=1	-1,663*	-	-	-
Bayesian	-0,022	0,062	-	-
K-Mean	-1,556	-1,501	-1,801*	-
K-Mean intercept	-1,097	-1,040	-1,227	0,652

Table 7. Diebold Mariano statistics for comparison of relative forecast quality for the different models. An asterisk indicates significance at 5 percent level.

model, which was stated by Diebold (1991). Because the volatility is a very fluctuating series, it doesn't seem reasonable to assume that the data generating process (DGP) follows a linear specification. Therefore, the true data generating model is not among the models that are being combined, and thus the method performs rather poorly. Both K-mean specifications offer very good results. Though none of their DM-statistics are significant, their negative values point towards a good performance compared to the other models. The best weighting scheme for the full sample has proven to be the regular inverse MSPE based weights. In Appendix G, we have included a graph of the moving Out-of-Sample R^2, showing that the mean and MSPE perform almost as well for all periods.

For the selection of the top models, the results are very different. Because of the very tight spread around the true value, these forecasts are all of high quality, resulting in good forecasts regardless of the weighting scheme that is selected. In table 8 a second comparison of all models is shown, but this time for the set of top 100 models.

Model	Mean top 100	MSPE top 100	Bayesian Average top 100	PCA top 100	K-Mean without intercept top 100
Mean, top 100	-	-	-	-	-
Median, top 100	0,231	0,467	1,408	-0,567	-1,381
MSPE, top 100	-0,201	-	-	-	-
Bayesian Average, top 100	2,423*	2,514*	-	-	-
PCA, top 100	0,688	0,679	-0,954	-	-
K-Mean without intercept, top 100	1,800*	1,705*	-0,489	0,794	-
K-Mean with intercept, top 100	1,724*	1,772*	0,119	-2,069*	1,033

Table 8. Diebold Mariano statistics for comparison of relative forecast quality for the different models over the sub selection of the top 100 forecasts. An asterisk indicates significance at 5 percent level.

A very different ranking emerges from this table than from the previous one, and there is no longer a model that performs better than the mean. Within the top 100 models, there is never a model that is always among this top during the full time period of 216 months. Different models perform well at different times, and the best model is very time variant. The mean always results in a good forecast, because all the forecasts lie evenly spread around the true value (see Figure D.2). The moving window of the MSPE causes the weights to only slowly adjust if a model is no longer the best, so it is probable that due to this not always the right models are selected. Where the effect of this is limited when looking at all models, this effect is more present in this context. The Bayesian Averaging model performs quite well, but is significantly outperformed by the more basic schemes. The K-Mean algorithm didn't result in better forecasts on the selection of forecasts as it did on the full sample.

We have seen that for the full sample of forecasts, the regular inverse MSPE based weights are the only weights that outperform the mean weights. The K-Mean algorithm also worked very well, but was not significantly better than the mean. For the selection of forecasts, there is no model that provides better forecasts than the mean.

4.3 Economical Evaluation

In this section the different forecasts will be evaluated economically, as described in Section 3.8. The results of the evaluation are shown in Table 4.6. In this table, the weights were performed on the full set of forecasts and the best specification of each model as found in the previous section is chosen as the model for this analysis.

Stocks	**Mean**	**STD**								
100% Market	17.28	54.27								
50% Market	11.45	25.16								
0% Market	3.81	1.66								
	Weights $\in [\mathbf{0}, \mathbf{1}]$					**Weights** $\in [-\mathbf{1}, \mathbf{2}]$				
Model	Mean	STD	Δ50	Δ100	$\Delta 0^5$	Mean	STD	Δ50	Δ100	$\Delta 0^5$
Real	11.45	20.76	1151	4480	22	12.24	22.98	910	4461	-1592
Mean	11.02	22.51	865	4361	42	11.28	22.98	814	4365	-2230
Median	11.13	22.90	813	4354	2	11.43	23.41	752	4358	-2227
MSPE	10.97	22.61	844	4351	2	11.23	24.09	791	4355	-2247
Bayesian	11.86	25.21	391	4313	21	12.33	26.32	4	4300	-2271
K-Mean no intercept	10.21	19.37	1185	4411	8	10.55	20.15	1133	4415	-2024
K-Mean intercept	9.89	18.74	1217	4403	8	10.14	19.32	1183	4406	-2028

Table 9. Economic evaluation of the stock return volatility forecasts of the constructed models and the benchmark model, performed on the full sample.
Note: **Delta50, Delta100 and Delta0 are the performance fees an investor is willing to pay extra to use the models instead of the standard strategies, displayed in basis points. Mean and STD are respectively the average and the standard deviation of the portfolio return. The 'real' model is the economic evaluation where the optimal weights were to be constructed with the real values for the variances. Two weightings schemes are used, where in the second weighting scheme short selling and lending is allowed.**

Table 9 shows that the 100% market strategy does have a very high mean return, but it also has a high standard deviation, which is not optimal for our considered investor. The risk-free strategy has a low standard deviation, but consequently a relatively low return. The optimal balance between risk and return would be accomplished by either the 50% or the 0% market strategy, dependent on our investors investment function. Investing based on the real value shows very promising results, as the return is equal and the standard deviation is smaller than that of the 50% market strategy. We can also see in the table that there is no model which would be better for this investor than the real volatilities, as based on the mean of the investment return. We do however see that the investor would pay a roughly equal performance for many of our weighting schemes, which is a good indication of the quality of these models. Specifically the K-Mean based weighting schemes show very promising results. Even though the mean of this strategy lies noticeably lower than that of the real volatility based strategy, the standard deviation is lower. For Delta 50, the performance fee that the investor is willing to pay is even larger than the fee that he would pay to use the real values. The

[5] For Delta 0, the risk aversion rate is set to γ=1. For the regular value of γ=6, the utility function of the 0% market buy-and-hold strategy did not coincide with the utility function of any model, indicating that our model performed insufficient.

Bayesian model shows the lowest rating, probably due to the high standard deviation of the return. In the second part of the table, where short selling and going long are allowed, we don't see results that vary much from the first part. The Delta 0 fees are all negative, indicating that we do not have any interesting information which would increase the investors' profit.

Stocks	Mean	STD								
100% Market	17.28	54.27								
50% Market	11.45	25.16								
0% Market	3.81	1.66								
	Weights $\in [\mathbf{0}, \mathbf{1}]$					**Weights** $\in [-\mathbf{1}, \mathbf{2}]$				
Model	Mean	STD	Δ50	Δ100	Δ0[6]	Mean	STD	Δ50	Δ100	Δ0[5]
Real	11.45	20.76	1151	4480	22	12.24	22.98	910	4461	-1592
Mean	11.04	22.69	838	4355	5	11.32	23.56	715	4342	-2226
Median	11.06	22.61	853	4360	5	11.39	23.63	709	4545	-2227
MSPE	10.86	22.57	840	4342	4	11.13	23.41	724	4330	-2224
PCA	9.99	18.59	1242	4418	22	10.2	18.98	1225	4425	-2215
Bayesian	11.31	24.17	593	4311	3	11.59	25.04	412	4295	-2258
K-Mean no intercept	10.21	19.37	1185	4411	8	10.55	20.15	1133	4415	-2024
K-Mean intercept	9.89	18.74	1217	4403	8	10.14	19.32	1183	4406	-2028

Table 10. Economic evaluation of the stock return volatility forecasts of the constructed models and the benchmark model, performed on the top 100 models. ***Note:*** **Delta50, Delta100 and Delta0 are the performance fees an investor is willing to pay extra to use the models instead of the standard strategies, displayed in basis points. Mean and STD are respectively the average and the standard deviation of the portfolio return. The 'real' model is the economic evaluation where the optimal weights were to be constructed with the real values for the variances. Two weightings schemes are used, where in the second weighting scheme short selling and lending is allowed.**

In the second table, the analysis is shown for the models on the top 100 forecasts. We see that the Bayesian model is still the worst model, but the PCA based weights show very good results. The PCA scheme shows performance fees that closely exceed those of the K-Mean schemes which showed to be the best when analyzing the full set of forecasts. A remarkable result is that, in the full sample, for the statistical evaluation the MSPE based weights performed best, but for the economical evaluation the K-Mean algorithm based weights prove to provide the best results. For the top 100 selection, the PCA analysis shows to work even better than the K-Mean scheme, even though in the statistical analysis it was shown that the mean proved to be the best model. Since the investor in this paper is rather risk-averse, the standard deviation is more important in the economic evaluation, causing the other models to provide better results.

[6] For Delta 0, the risk aversion rate is set to γ=1. For the regular value of γ=6, the utility function of the 0% market buy-and-hold strategy did not coincide with the utility function of any model, indicating that our model performed insufficient.

5. Conclusion

This paper examined the effects of combining a set of one-step-ahead forecasts from linear regression models using different weighting schemes. The weights were used both on the full set of forecasts as well as a selection of the top 100, to find a weighting scheme that could outperform the simple average of these forecasts, which has proven to be a very solid benchmark in previous research.

Different conclusions can be drawn from this research. For the full sample where there are many forecasts to be combined, there are big opportunities for more advanced weighting schemes. The inverse MSPE based weights perform very well compared to the mean in the statistical analysis. The Bayesian averaging scheme did not prove as valuable as expected, probably due to the violation of the assumption about the DGP in this dataset. Both in the statistical as well as in the economical evaluation, this model did not give valuable results. The results for the K-mean algorithm are very promising, and results indicate that this scheme could be a good candidate for further research. Even though the specification of the K-Mean algorithm was very basic, it showed good results in both the statistical as well as the economical evaluation.

For the selection of the top models however, advanced weighting schemes are fairly redundant when reviewed in terms of statistical properties. Because of the good quality of the selected forecasts to combine, there is fairly little space for improvements regarding the weights. For an investor however, there are weighting schemes that provide him with more profit than the mean of the forecasts could, so performance depends on the area of usage.

We conclude that the use of more advanced weighting schemes is most useful when the sample of forecasts is large, because here the quality difference is very big. For small sets of forecasts, of more or less equal quality, the usefulness of more advanced methods depends on the use of the forecasts.

References

Aiolfi., M., Timmermann, A., 2006. Persistence in forecasting performance and conditional combination strategies. Journal of Econometrics 135, 31-53.

Baridam., B.B., 2012. More work on K -Means Clustering Algorithm: The Dimensionality Problem. *International Journal of Computer Applications* 44(2):23-30

Bates, J.M., Granger, C.M.W., 1969. The combination of forecasts. *Operations Research Quarterly 20*, 451-468.

Buckland, S.T., Burnhamn, K.P., Augustin, N.H., 1997. Model selection: An integral part of inference. Biometrics 53, 603-618.

Cakmakli, C. and D. Van Dijk., 2013, Getting the Most out of Macroeconomic Information for Predicting Excess Stock Returns, Tinbergen Institute Discussion Paper 2010-115/4.

Christiansen, C., M. Schmeling and A. Schrimpf., 2012. A comprehensive look at financial volatility prediction by economic variables, *Journal of Applied Econometrics* 27, 956-977.

Diebold, F.X., 1991. A Note on Bayesian Forecast Combination Procedures. In A. Westlund and P. Hackl (eds.) Economic Structural Change: Analysis and Forecasting, Springer-Verlag, 225-232.

Gupta, S.,Wilton, P.C. 1987. Combination of forecasts: An extension. Management Science 33, 356–372.

Holtrop, N., W. Kers, F.C.A. Mourer and M.G.T. Verkuijlen., 2014. Volatility's Next Top Driver: Forecasting volatility with the use of macroeconomic and financial variables. (To be published).

Hsiao, C., & Wan, S., 2014. Is there an Optimal Forecast Combination? *Journal of Econometrics*, *178*, 294-309.

Marcellino, M., 2004. Forecast pooling for short time series of macroeconomic variables. Oxford Bulletin of Economic and Statistics 66, 91–112.

Palm, F.C., Zellner, A., 1992. To combine or not to combine? Issues of combining forecasts. Journal of Forecasting 11, 687–701.

Stock, J.H., Watson, M., 2001. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: Engle, R.F., White, H. (Eds.), Festschrift in Honour of Clive Granger. Cambridge University Press, Cambridge, pp. 1–44

Timmermann, Allan G.2005, Forecast Combinations. CEPR Discussion Paper No. 5361. Available at SSRN: http://ssrn.com/abstract=878546

Timmermann, A., 2006, Forecast Combinations, *Handbook of Economic Forecasting* 1, 135-196.

West, K. D., H. J. Edison and D. Cho., 1993, A Utility-Based Comparison of Some Models of Exchange Rate Volatility, *Journal of International Economics*, 35, pp. 23-45.

Appendices

Appendix A : Statistical properties of regressors

This following table displays the basic statistical properties of the variables that were used to forecast the volatility, extracted from the paper of Holtrop(2014)

Variable	Abbrev.	Mean	Std.	Skew.	Kurt.	AC(1)
A. Equity Market Variables and Risk Factors						
1 Dividend Price Ratio (Log)*	D-P	0.28	4.58	0.77	6.44	0.06
2 Earnings Price Ratio (Log)	E-P	-3.02	0.43	-1.31	6.49	0.98
3 US Market Excess Return	MKT	0.59	4.57	-0.91	5.77	0.10
4 Size Factor	SMB	0.12	3.23	0.81	11.44	-0.03
5 Value Factor	HML	0.35	3.15	0.05	5.54	0.14
6 Short Term Reversal Factor	STR	0.37	3.44	0.17	8.34	-0.02
7 S&P500 Turnover	TURN	0.01	0.16	-0.07	3.38	-0.51
8 Return MSCI World	MSCI	0.73	4.26	-1.20	6.44	0.13
B. Interest Rates, Spreads and Bond Market Factors						
9 T-Bill Rate (Level)*	T-B	-0.23	2.32	0.95	5.12	0.48
10 Rel. T-Bill Rate	RTB	-0.18	0.86	-0.30	2.85	0.95
11 Long Term Bond Return	LTR	0.81	2.97	0.20	4.78	0.02
12 Rel. Bond Rate	RBR	-0.18	0.63	-0.36	4.49	0.87
13 Term Spread*	T-S	-0.01	33.77	0.34	3.67	0.08
14 Cochrane Piazzesi Factor	C-P	1.22	1.56	0.41	3.34	0.90
C. FX Variables and Risk Factors						
15 Dollar Risk Factor	DOL	0.12	2.19	-0.34	4.02	0.12
16 Carry Trade Factor	C-T	0.05	2.58	-0.69	4.38	0.18
17 Average Forward Discount	AFD	0.18	0.19	0.87	7.83	0.75
D. Liquidity and Credit Risk Variables						
18 Default Spread	DEF	0.11	0.43	2.48	12.3	0.94
19 FX Average Bid-ask Spread	BAS	1.43	5.00	1.92	7.46	0.88
20 Pastor-Stambaugh Liquidity Factor	PS	-0.28	6.83	-1.76	10.49	0.00
21 TED Spread	TED	0.07	0.00	1.78	8.67	0.81
E. Macroeconomic Variables						
22 Inflation Rate, Monthly	INFM	0.24	0.31	-1.38	11.31	0.47
23 Inflation Rate, Yearly	INFA	2.91	1.26	-0.48	4.41	0.95
24 Industrial Production Growth, Monthly	IPM	0.20	0.66	-1.32	10.18	0.23
25 Industrial Production Growth, Yearly*	IPA	0.27	9.52	0.29	6.96	0.28
26 Housing Starts	H-S	-2.20	24.9	-0.04	4.52	0.79
27 M1 Growth, Monthly	M1M	0.40	0.79	1.51	13.79	0.18
28 M1 Growth, Yearly	M1A	4.81	4.98	0.29	2.31	0.98
29 Orders, Monthly	ORDM	0.11	1.78	0.00	3.15	-0.19
30 Orders, Yearly	ORDA	1.20	6.93	-1.51	8.49	0.93
31 Return CRB Spot	CRB	0.25	2.74	-1.76	17.62	0.25
32 Capacity Utilization	CAP	0.02	0.66	-1.07	8.95	0.25
33 Employment Growth	EMPL	0.11	0.19	-0.37	7.40	0.65
34 Consumer Sentiment	SENT	0.01	4.70	0.07	5.66	0.00
35 Consumer Confidence	CONF	0.02	8.25	-0.29	9.94	0.07

36 Diffusion Index	DIFF	8.68	16.91	-0.64	3.57	0.83
37 Chicago PM Business Barometer	PMBB	55.15	7.33	-0.37	3.37	0.88
38 ISM PMI	PMI	52.08	5.35	-0.39	3.77	0.93

Table A.1 . Basic statistical properties of all the regressors.

Note: **The table shows the summary statistics for the 38 macro-economical and financial predictive variables. The reported statistics include the mean, standard deviation (Std.), Skewness (Skew.), Kurtosis (Kurt.), as well as the first order autocorrelation coefficient (AC(1)). An asterisk (*) denotes that the variable is changed from Christianssen et al. (2012), corrected for a unit root.**

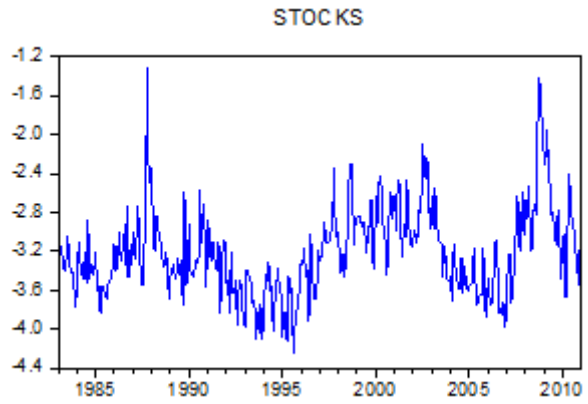


Table A.2 . Graph of the realized volatility of stocks over the entire horizon

Realized Volatility Stocks	
Mean	-3.21
Standard dev.	0.45
Skewness	0.81
Kurtosis	4.44
JB P-value	0.00
AC(1)	0.71

Table A.3 . Basic statistical properties of the realized volatility of stocks

Appendix B : Histograms of forecast error distribution

Below are three histograms displaying the distribution of the forecast errors for all the 2^{18}=262.144 forecasts. The histograms are created for the first forecast at time T_1+1, the 100th forecast and the 200th forecast. For all three histograms, we can clearly conclude that the forecast errors are not standard normal distributed.

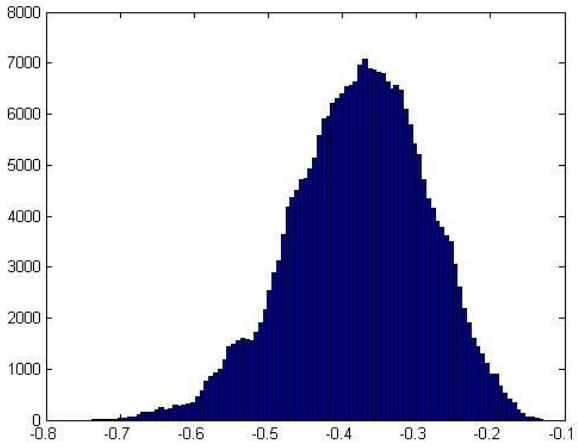


Figure B1: Histogram of forecast errors of all models at time t=T1+1.

***note**: The y axis refers to the amount of models, the x-axis is the size of the forecast error*

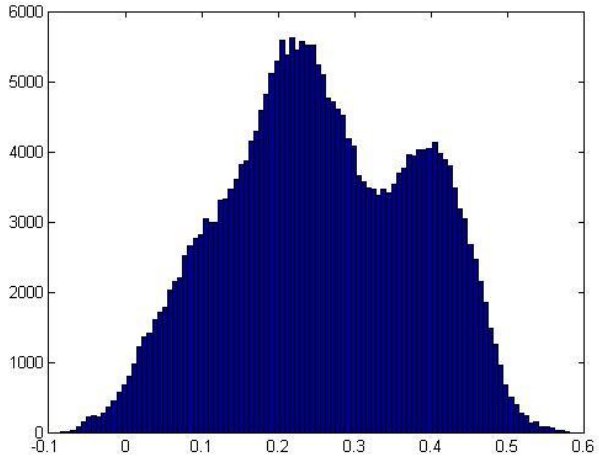


Figure B2: Histogram of forecast errors of all models at time t=T1+100.

***note**: The y axis refers to the amount of models, the x-axis is the size of the forecast error*

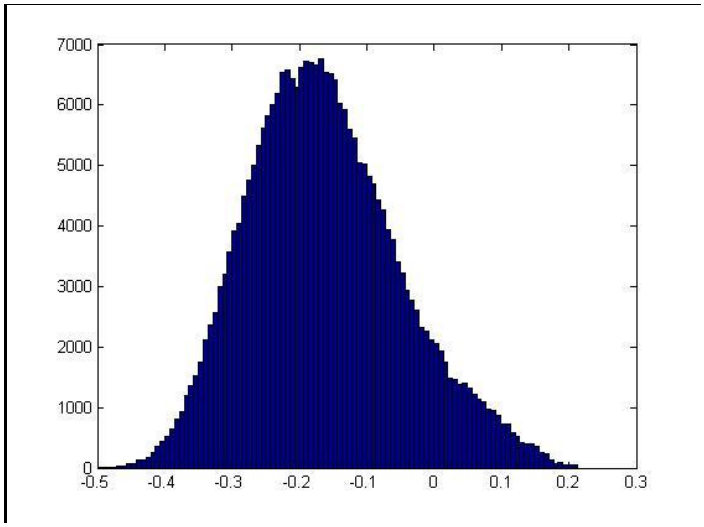


Figure B3: Histogram of forecast errors of all models at time t=T1+200.

***note**: The y axis refers to the amount of models, the x-axis is the size of the forecast error*

Appendix C : Derivations of weight properties

Appendix C.1 : Derivation of the optimal weights for a combination of forecasts

Say that we have two sets of forecasts y_1 and y_2 with corresponding forecast errors e_1 and e_2, which we want to combine by assigning weights w and $(1 - w)$. The forecast error for their combination than becomes $e' = we_1 + (1 - w)e_2$, with mean zero and variance equal to

$$\sigma^2 = w^2\sigma_1^2 + (1 - w)^2 \sigma_2^2 + 2w(1 - w)\sigma_{12}$$

We can then derive the optimal weights $\bar{w}$ by differentiating the models variance with respect to *w*, which gives $w^* = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$

The model variance of these optimal weights can be obtained by substituting w with w* in the model variance equation, resulting in a model variance equal to

$$\sigma_{opt}^2(w^*) = \frac{\sigma_1^2\sigma_1^2(1-{\rho_{12}}^2)}{\sigma_1^2+\sigma_2^2-2\sigma_1\sigma_2\rho_{12}}$$

This expressions is smaller than the minimum of σ_1^2 and σ_2^2 , and diversification is useful as long as non of the following cases occur: σ_1 or σ_2 equal to zero; $\sigma_1 = \sigma_2$ and $\rho_{12} = 1$ or $\rho_{12} = \frac{\sigma_1}{\sigma_2}$

Appendix C.2 : Derivation of the model variance ratio of mean vs optimal weights

For an equal weighted combination of the two forecasts, $y = \frac{1}{2}y_1 + \frac{1}{2}y_2$, the model variance now becomes

$$\sigma_{mean}^2 = \frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{2}\sigma_1\sigma_2\rho_{12}$$

Dividing this on the optimal model variance yields, after some rearranging

$$\frac{\sigma_{mean}^2}{\sigma_{opt}^2} = \frac{(\sigma_1^2 + \sigma_2^2)^2 - 4{\sigma_{12}}^2}{4\sigma_1^2\sigma_2^2(1 - \rho_{12}^2)}$$

Appendix C.3 : Derivation of the model variance ratio of MSPE vs optimal weights

For a weight based on the MSPE, the forecast becomes $y = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}y_1 + \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}y_2$, with corresponding model variance equal to

$$\sigma_{MSPE}^2 = \frac{\sigma_1^2\sigma_2^2(\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\rho_{12})}{(\sigma_1^2+\sigma_2^2)^2}$$

Dividing this on the optimal model variance yields, after some rearranging

$$\frac{\sigma_{MSPE}^2}{\sigma_{opt}^2} = \left(\frac{1}{1 - \rho_{12}^2}\right)\left(1 - \left(\frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2}\right)^2\right)$$

Appendix D: Plots of the forecasted values against the true value.

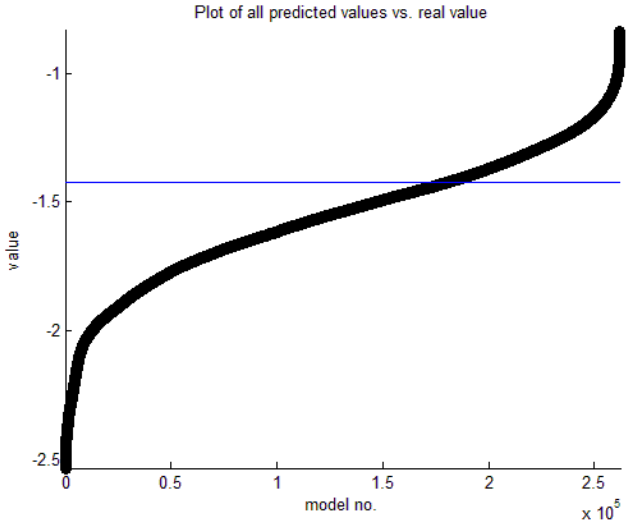


Figure D.1 : Display of all the 2^{18} = 262.144 forecasts against the true value.

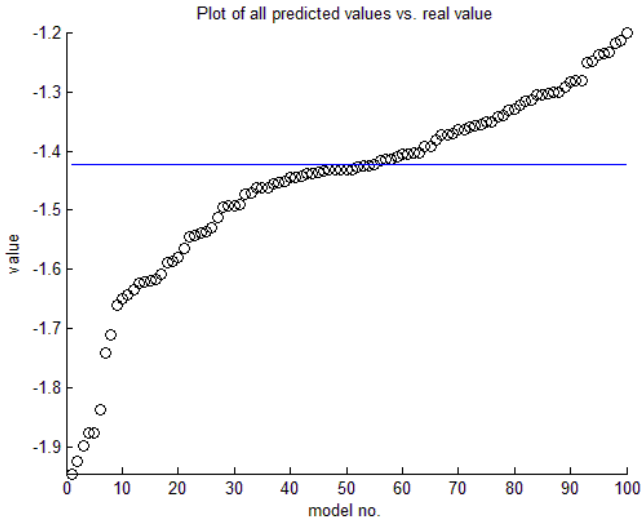


Figure D.2 : Display of the selection of 100 forecasts against the true value.

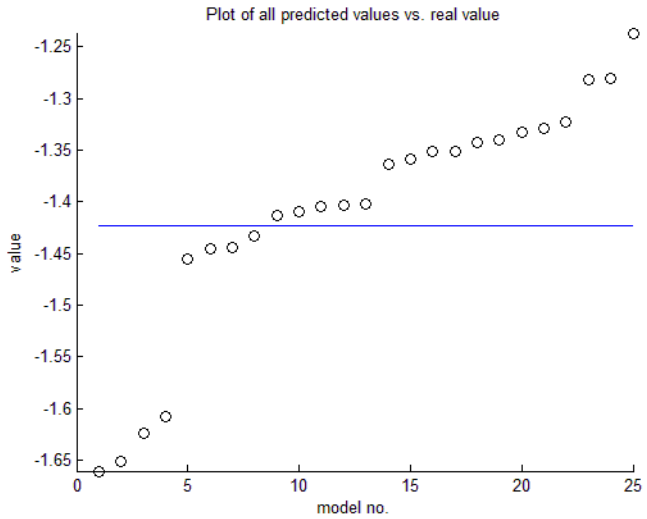


Figure D.3 : Display of the selection of 25 forecasts against the true value.

Appendix E : K-Mean-Clustering algorithm

Appendix E.1 . Groups for the K-Mean-Clustering

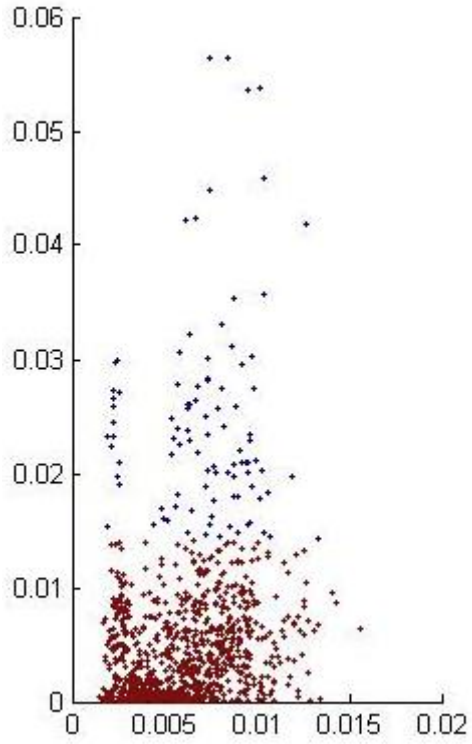


Figure E.1 . Plot of the group division as found by the K-mean-Clustering algorithm. ***Note: The y-axis contains the average squared forecast error over the last 4 periods, and the x-axis contains the squared forecast error of the current period.***

The groups have been chosen based on current performance with respect to performance over the past 4 periods. Models that did well over the last 4 periods(having a low y-value) and still do well(having a low x-value), are chosen by this algorithm as members of the first group. This interpretation can be extended in a very straightforward way if more clusters need to be chosen.

Appendix E.2 . Plots of the regression coefficients over time

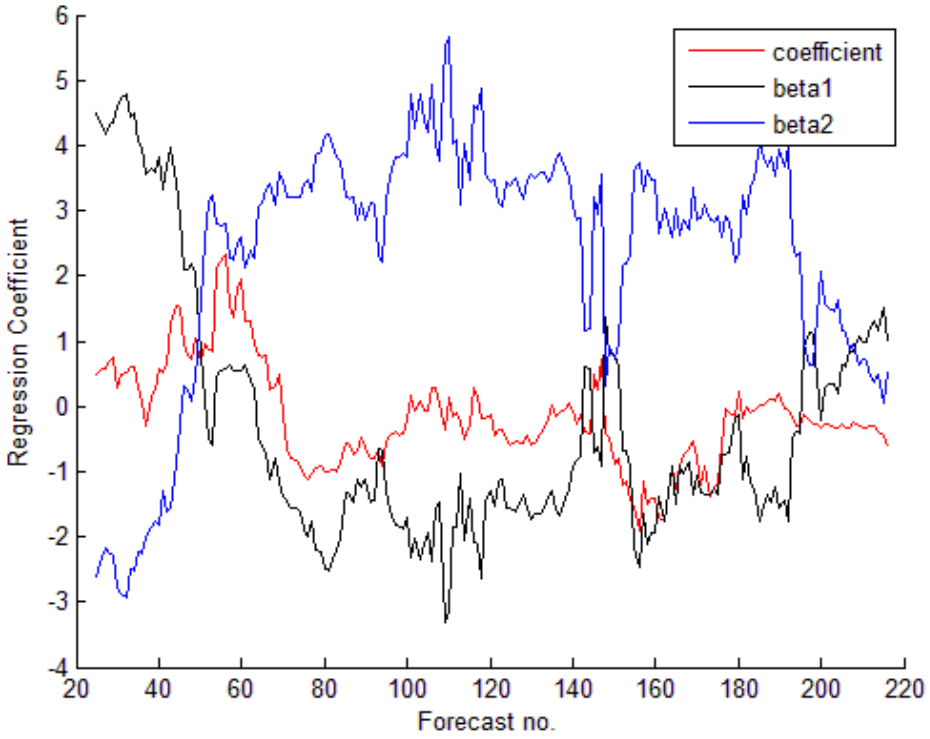


Figure E.2 : Plot of the regression coefficients over time for the regression with the intercept.

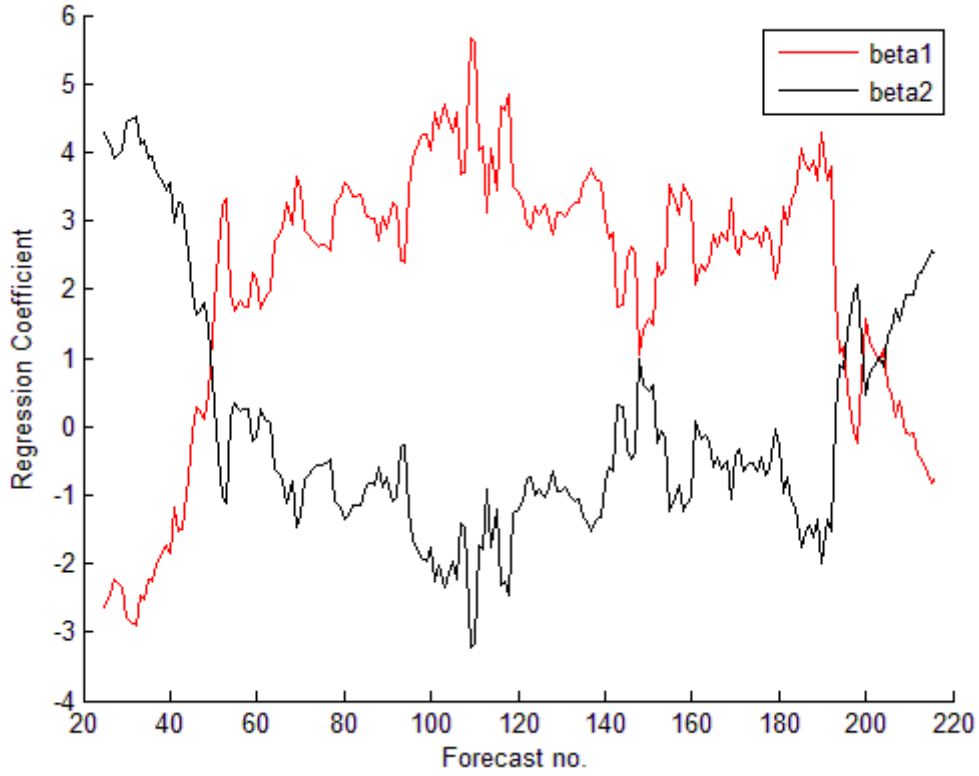


Figure E.3 : Plot of the regression coefficients over time for the regression without the intercept

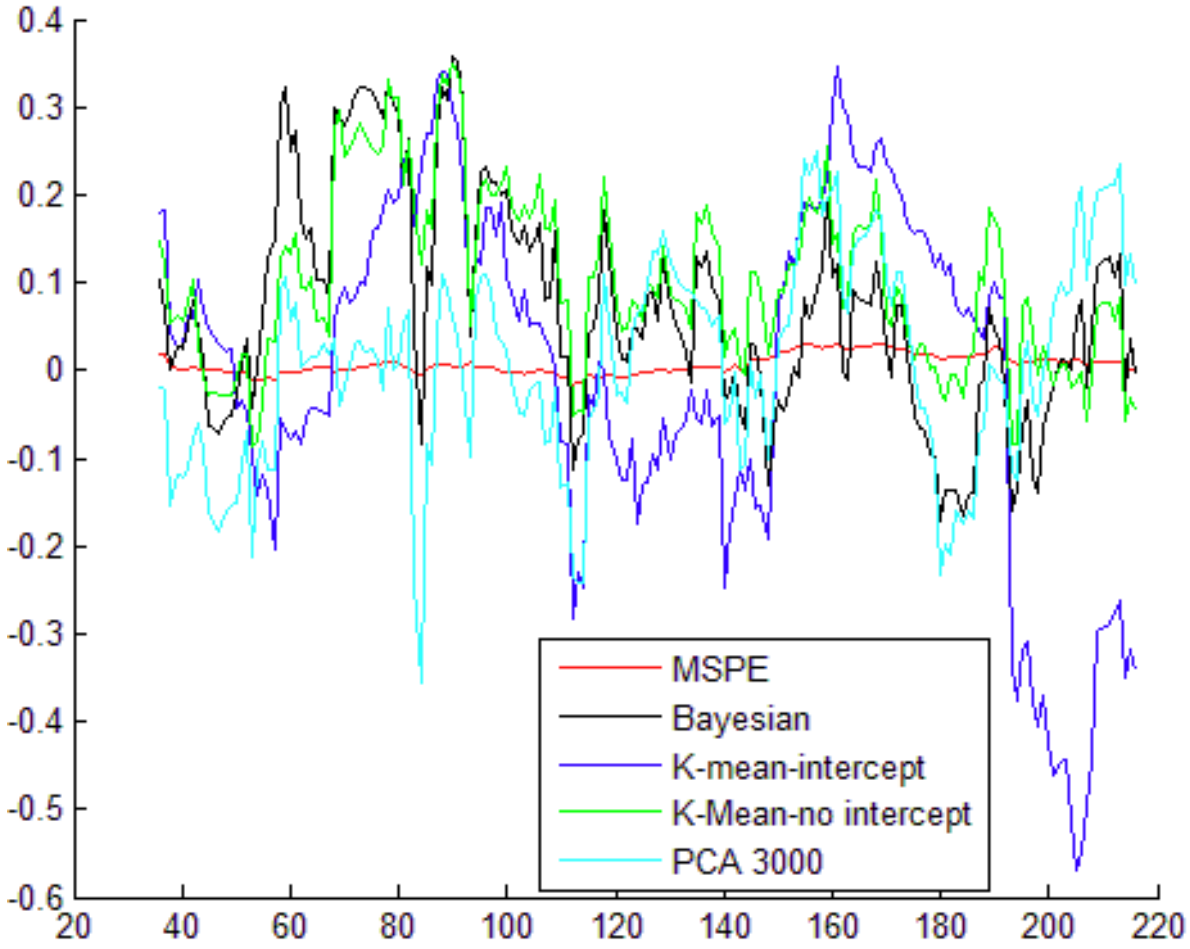
Appendix F . DM-statistics for bigger selections of forecasts

Model	Mean	Mean top25	Mean top100	Mean top500	Mean top1000
Mean					
Mean top25	-1,757*				
Mean top100	-2,164*	-1,711*			
Mean top200	-2,220*	-0,998	0,258		
Mean top500	-2,564*	-1,205	-0,469	-1,267	
Mean top1000	-2,840*	-1,179	-0,582	-1,043	-0,717

Table F.1 . Forecast quality comparison by means of Diebold Mariano statistics. ***Note: Values smaller than the critical value of -1,96 indicate that the model in the row performs better, and values larger than 1,96 indicate that the model in the column performs better. Critical values are at the 5 percent level.***

We can see that all the forecasts from selections outperform the forecast of the mean on the full sample of forecasts, but there is not a single model which outperforms the mean on the top 100. We can safely conclude that this is a good top selection to use.

Appendix G . Moving Out-of-Sample R^2



Tabel G.1 . Graph over the moving out of sample R^2, moving window length 24 months, with the mean as benchmark and all the forecasts as sample. ***Note: A positive value implies outperformance of the mean, while a negative value implies that the model performs worse than the benchmark model.***

We can clearly see the good all-round performance of the MSPE based scheme, as well as the K-Mean no intercept scheme. Both have a positive OOS R^2 most of the time, indicating a better performance than the mean. The other schemes performances fluctuate very much, with more negative than positive peaks, indicating their worse performance.

Appendix H . Transformation of volatility to realized variance

This derivation has been extracted from the report by Holtrop(2014).

In this paper the realized volatility has been forecasted. Transforming the forecasts for the realized volatility into those of realized variance is done in the following way. The realized volatility is given by

$$RV_{i;t} = ln\sqrt{\sum_{\tau=1}^{M_t} r^2{}_{i;t;\tau}} \qquad t = 1, \dots, T$$

The realized volatility ($RV_{i;t}$) is the natural logarithm of the square root of the realized variance, the realized variance is set as X.

$$X_{i;t} = \sum_{\tau=1}^{M_t} r^2{}_{i;t;\tau} \qquad t = 1, \dots, T$$

In this research forecasts are made for the realized volatility (RV). In order to transform these results into the realized variance X, it is needed to determine the expected value of X, given the distribution of the RV. Suppose that the realized volatility has a normal distribution with mean μ and variance σ^2:

$$RV_{i;t} \sim N(\mu, \sigma^2)$$

Rewriting the RV gives:

$$RV_{i;t} = ln\sqrt{X_{i;t}} = \frac{1}{2} ln\, X_{i;t} \sim N(\mu, \sigma^2)$$

One can say that:

$$ln\ X_{i;t} \sim N(2\mu, 4\sigma^2)$$

$$since\ \ln(U) \sim N(\mu, \sigma^2)\ then\ E(U) = e^{\mu + \frac{1}{2}\sigma^2}$$

This results into the expected value of the realized variance X:

$$E(X_{i;t}) = e^{2\mu + 2\sigma^2}$$