International Journal of Forecasting 31 (2015) 943–951

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems

Andreas Graefe [a,*], Helmut Küchenhoff [b], Veronika Stierle [b], Bernhard Riedl [b]

[a] Department of Communication Science and Media Research, LMU Munich, Germany

[b] Institute of Statistics, LMU Munich, Germany

ARTICLE INFO

Keywords:
Bayesian analysis
Combining forecasts
Economic forecasting
Election forecasting
Equal weights

ABSTRACT

We compare the accuracies of simple unweighted averages and Ensemble Bayesian Model Averaging (EBMA) for combining forecasts in the social sciences. A review of prior studies from the domain of economic forecasting finds that the simple average was more accurate than EBMA in four studies out of five. On average, the error of EBMA was 5% higher than that of the simple average. A reanalysis and extension of a published study provides further evidence for US presidential election forecasting. The error of EBMA was 33% higher than the corresponding error of the simple average. Simple averages are easy both to describe and to understand, and thus are easy to use. In addition, simple averages provide accurate forecasts in many settings. Researchers who are developing new approaches to combining forecasts need to compare the accuracy of their method to this widely established benchmark. Forecasting practitioners should favor simple averages over more complex methods unless there is strong evidence in support of differential weights.

1. Introduction

Ensemble Bayesian Model Averaging (EBMA) is a relatively new approach to combining forecasts that emerged from the data-heavy domain of weather forecasting (Raftery, Gneiting, Balabdaoui, & Polakowski, 2005). EBMA calculates a weighted average of forecasts, where the weights are based on the past performance and uniqueness of each component forecast. Montgomery, Hollenbach, and Ward (2012), hereafter MHW, test the performance of EBMA for three subject areas (insurgencies, US presidential elections, and US Supreme Court decisions) within the domain of political forecasting. MHW find that the EBMA forecasts are more accurate than the individual component forecasts in each case and propose that the method be used widely for forecasting social science problems.

Combined forecasts are often more accurate than the individual component forecasts. In a meta-analysis of thirty studies, the simple unweighted average of multiple forecasts had errors 12% lower than those of the typical individual forecast (Armstrong, 2001). In addition, the average forecasts were often more accurate than the most accurate component forecast. Armstrong's analysis also indicated that the gains in accuracy from combining are expected to be highest when five or more forecasts can be obtained, when the forecasts draw upon different validated methods and data, and when there is uncertainty as to which forecast is most accurate.

Graefe, Armstrong, Jones, and Cuzán (2014) analyzed the gains from combining forecasts when forecasting US presidential elections by averaging forecasts within and across four different methods: polls, prediction markets, expert judgment, and quantitative models. This approach, the results of which are published at PollyVote.com, yielded large gains in accuracy, much larger than the 12%

* Corresponding author.
E-mail address: a.graefe@lmu.de (A. Graefe).

http://dx.doi.org/10.1016/j.ijforecast.2014.12.001
0169-2070/

error reduction previously estimated by Armstrong (2001). Across the six elections from 1992 to 2012, the combined PollyVote forecast was more accurate than any one of the component methods; on average, the error reductions ranged from 16% (compared to prediction markets) to 59% (compared to polls).

We applaud MHW for their promotion of the use of combined forecasts. By focusing on combining in their study, they raised awareness of a powerful method that is still underutilized in both research and practice in many fields. In most situations, people will make better predictions by combining forecasts from several sources, rather than by relying on a single source.

However, there is a pitfall. EBMA involves a level of complexity that is often unnecessary when combining forecasts. Innumerable studies on combining have shown that simple combining procedures, such as calculating unweighted averages of forecasts, provide a benchmark that is hard for more complex approaches to beat (Clemen, 1989). MHW do not compare EBMA to this widely accepted benchmark adequately, nor do they discuss the conditions under which EBMA is expected to be useful.

The present study summarizes prior evidence on the relative performances of variants of EBMA and the simple average for combining economic forecasts. We then provide new evidence by reanalyzing and extending MHW's analysis of US presidential election forecasts. We find that EBMA contributes little to the accuracy of simple averages when combining forecasts for social science problems.

2. The issue of weights in combining forecasts

A widespread concern when combining forecasts is the question of how best to weight the components, and many scholars have proposed different methods for doing so. However, an early review of more than two hundred published papers from the fields of forecasting, psychology, statistics, and management concluded that the question of *how* to combine forecasts does not seem to be critical to the forecast accuracy. In fact, it was found that the simple average (i.e., assigning equal weights to components) often provides more accurate forecasts than complex approaches to estimating "optimal" combining procedures (Clemen, 1989).

The empirical research since then has repeatedly confirmed these findings. A recent example is the large-scale comparison of simple averages and various sophisticated approaches to the combination of economic forecasts from the European Central Bank's Survey of Professional Forecasters by (Genre, Kenny, Meyler, & Timmermann, 2013). The sophisticated methods included combinations based on principal components, trimmed means, performance-based weighting, optimal least squared estimates, and Bayesian shrinkage. The performances of these methods varied over time, across target variables, and across time horizons. Moreover, any predictive gains relative to an equal weighting of forecasts were shown to be due probably to chance. The authors therefore concluded that there is only a modest case for considering combinations other than equal weighting as a means of summarizing the survey replies better.

An analysis of the relative performances of several combining procedures based on a seven-country data set of economic forecasts made over the period from 1959 to 1999 provided similar results (Stock & Watson, 2004). Simple averages of all available forecasts provided more accurate predictions than sophisticated combination methods, which relied heavily on historical performances for weighting the component forecasts.

One reason for the strong performance of equal weights is the fact that the accuracy of the component forecasts varies over time and depends strongly on external effects. For example, in the study by Stock and Watson (2004), the accuracies of individual forecasts were influenced heavily by economic shocks and political events. Therefore, a good performance in one year or country did not predict a good performance in another. In such a situation, differential weights are of course of limited value.

Another possible explanation is estimation error in the differential weights. Smith and Wallis (2009) investigated this question by conducting a Monte Carlo simulation of combinations of two forecasts, and reappraising a published study using different combinations of multiple forecasts of US output growth. They concluded that the simple average will be more accurate than estimated "optimal" weights if two conditions are met: (1) the combination is based on a large number of individual forecasts and (2) the optimal weights are close to equality. The reason for this is that, in such situations, each forecast has a small weight, and the simple average provides an efficient trade-off against the error that arises from the estimation of weights.[1]

3. Ensemble Bayesian model averaging

Researchers' quest for an optimal solution to forecast combination continues in many fields. EBMA is a relatively new approach to the differential weighting of forecasts, and has become popular in the data-heavy domain of weather forecasting (Raftery et al., 2005). Simply put, EBMA calculates a probabilistic forecast distribution as a weighted average of component forecasts. The weights of the component forecasts are based on a statistical analysis of each individual component's past performance and the

[1] This body of literature is related closely to that on the relative performances of using equal and differential weights for the predictors in linear models. In many fields, the common procedure when developing linear models for predicting any kind of target variable is to identify a subset of most important predictors and to estimate the weights that provide the best possible solution *for a given sample*. The resulting "optimally" weighted linear composite is then used *when predicting new data*. While this approach is useful in situations with large and reliable datasets and few predictor variables, a large body of empirical and analytical evidence shows that the estimation of variable weights from the data is of little, if any, value in situations with small and noisy datasets and large numbers of predictor variables. Graefe (in press) provides an overview of this literature and demonstrates the gains from weighting the predictors equally in linear models for the task of US presidential election forecasting. Across the ten elections from 1976 to 2012, equally weighting the predictors used by established regression models reduced the forecast error for six of the nine models. On average across the ten elections and nine models, equal weights reduced the error of the original regression models by 5%.

uniqueness of the forecasts. That is, EBMA assigns higher weights to components that were more accurate historically and whose forecasts differ from the other forecasts in the ensemble. For a detailed description of the EBMA mechanism, see MHW or Raftery et al. (2005).

At first glance, two features of EBMA are theoretically appealing. First, the estimation of component weights based on a statistical a-priori analysis may ensure objectivity, as the mechanical approach prevents analysts from weighting the forecasts in a way that suits their biases. Second, the idea of assigning higher weights to forecasts that provide unique information is consistent with the recommended principle of combining forecasts that incorporate diverse information (Armstrong, 2001). The underlying idea is that the systematic and random errors of individual forecasts are more likely to cancel out in the aggregate if the individual forecasts draw on different information, and thus, are likely to be uncorrelated. However, despite its theoretical appeal, the mechanism EBMA uses to estimate the weights of the component forecasts raises concerns.

3.1. Estimation of weights in EBMA

A first concern with EBMA relates to the weighting of component forecasts based on their past performances. The supposition that a method's past performance is an indicator of its future performance may be plausible and appeal to common sense. However, as was noted in Section 2 – and as will be shown in Section 4.2.1 for US presidential election forecasts – it is often contradicted by empirical evidence. In many real-world forecasting problems, a model's past performance is unrelated to its predictive performance.

A second concern relates to the weighting of component forecasts based on their estimated uniqueness. EBMA assigns lower weights to components that provide forecasts that are similar to those from other components. This approach appears somewhat problematic, because it cannot ensure that unique information is included. In fact, it might even *prevent* the combined forecast from incorporating unique information. Imagine two forecasts that use different approaches and are based on different information (i.e., variables), but happen to provide similar predictions in the training data. In such a situation, EBMA would decrease the weight of (or even exclude) one of the forecasts. That is, EBMA would ignore information, thus probably harming the accuracy of the combined forecast when predicting new data. In other words, EBMA cannot account for unique information, as the method only considers *outputs* (i.e., the forecasts), ignoring the *inputs* (i.e., the underlying information).

A third concern relates to the data requirements of EBMA. Whereas standard forecasting applications require the dataset to be split into two parts (i.e., a training set to calibrate the forecasting model and a test set to assess the model's predictive accuracy), standard EBMA implementations require the dataset to be divided into three parts: (1) a training set, for fitting the component models, (2) a validation set, for estimating the components' prediction errors outside the sample used by the training set, in order to determine the components' weights when calculating the combined out-of-sample forecasts, and (3) a test set, for assessing the predictive accuracy out-of-sample. A typical recommendation in the literature is to split the dataset to give 50% for training, and 25% each for validation and testing (Hastie, Tibshirani, & Friedman, 2001, p. 196). Due to data limitations, such an approach is often problematic in the social sciences.[2]

3.2. Empirical evidence on the relative accuracies of Bayesian model averaging and simple averages

Raftery et al. (2005) provide evidence on the relative accuracies of EBMA and the simple average in the domain of weather forecasting. The study combines forecasts from five component models in order to generate 48-hour forecasts of temperature and sea level pressure over a six-month period. In a first step, the authors analyze the data to determine the optimal length of the training period for their particular prediction problem, which they estimate to be 25 days. Then, they use this training period to calibrate the component models and to estimate the EBMA parameters, which are finally used for calculating the combined EBMA forecast. On average across the two prediction tasks, the mean absolute error (MAE) of the EBMA forecasts was 6% lower than that of the simple average.

In the social sciences, evidence on the relative performances of EBMA and the simple average comes primarily from the domain of economic forecasting. Table 1 summarizes the results from five studies that report both the errors of variants of EBMA and those of a simple average of component forecasts. In each case, we calculated the relative error across all forecasts and all time horizons that were reported in the original studies. Table 1 shows the results. Values above one mean that EBMA was less accurate than the simple average; values below one mean that EBMA was more accurate. For example, the value in the second row shows that the ratio of the EBMA error to that for the simple average in the study by Stock and Watson (2005) was 1.14. That is, the error for the EBMA forecasts was 14% higher than that for the simple average.

Stock and Watson (2005) analyzed the relative accuracies of the simple average and EBMA for forecasting nine monthly economic time series. The forecasts were constructed by combining 131 individual models. The initial parameters of any models that require historical data were estimated based on data from January 1959 to June 1974. The accuracies of their out-of-sample predictions were then analyzed from July 1974 to December 2003, across four time horizons (1, 3, 6, and 12 months ahead). Across the full out-of-sample period and the four time horizons,

[2] For example, Ulfelder (2013) struggled with the question of how to split the data when he tried to use EBMA for forecasting mass-killings. Because of the small number of mass killings in more recent years and the uncertainty about their causes and incidents over time, he could not find any reasonable way to divide his dataset into training, validation, and test sets. In the end, Ulfelder decided to go with the evidence-based approach to forecast combination, and calculated unweighted averages.

Table 1
Relative accuracies of EBMA variants and the simple average for economic forecasting.

Target criterion	Error measure	Error ratio	Source
Output, prices, interest rates	Root-MSE	1.14	Stock and Watson (2005), Table 4
Output, prices, interest rates	Root-MSE	1.09	Clark and McCracken (2010), Tables III, IV and V
House prices	Relative-MSE (compared to a constant forecast)	1.03	Bork and Møller (2015), Table 1
Exchange rates	Relative-MSE (compared to driftless random walk)	1.03	Wright (2008), Tables 1, 2 and 8
Inflation	Root-MSE	0.97	Wright (2009), Table 2
	Mean	**1.05**	

the simple average was more accurate than each of four EBMA variants, whose performances varied depending on prior assumptions about the model parameters. The forecast errors of the EBMA variants were 7%–23% higher than that of the simple average, or 14% on average.

Clark and McCracken (2010) analyzed the relative accuracies of the simple average and EBMA for quarterly forecasts of US output, prices, and interest rates based on combinations of (subsamples of) 50 univariate models. Any models that require parameter estimation were fitted initially on data from 1965:Q4 to 1969:Q4 and then successively updated. The accuracies of the out-of-sample forecasts across four time horizons (0, 1, 4, and 8 quarters ahead) were then evaluated from 1970:Q1 to 2005. Based on the reported results, we calculated average forecast errors across all target variables, all samples, and all time horizons (144 observations). We found that the forecast errors of the three EBMA variants were 5%, 9%, and 14% (or 9% on average) higher than the corresponding error of the simple average.

Bork and Møller (2015) provide evidence for the forecasting of US house prices at the state level. The models were initially fitted using data from 1976:Q2 to 1994, and then successively updated. The accuracies of the out-of-sample forecasts were analyzed from 1995:Q1 to 2012. Across all 50 states and the full forecast horizon, the forecast error of EBMA was 3% higher than that of a simple average of regression models.

Wright (2008) analyzed the accuracies of 12 EBMA variants for monthly and quarterly exchange rate forecasts of four currencies. The performances of the EBMA variants depended on the choice of the shrinkage parameter, with more shrinkage yielding better performances. Across all forecasts and time horizons, the simple average had a lower error than six of the 12 EBMA variants. On average across the 12 EBMA variants, the forecast error of EBMA was 3% higher than that of the simple average. In a similar study, Wright (2009) compared the relative accuracies of the simple average and EBMA for quarterly US inflation forecasts. The combined forecasts were based on 107 individual models and were calculated for one to eight quarters ahead. Across the full study period from 1971 to 2006 and for all time horizons, the forecast error of EBMA was 3% lower than that of the simple average.

In summary, the empirical evidence on the relative accuracies of forecast combination using EBMA and simple averages is mixed. While EBMA yielded small accuracy gains in the domain of weather forecasting, four out of five studies from the domain of economic forecasting find that EBMA provided predictions that were less accurate than those of simple averages. On average across the four economic forecasting studies, the error of EBMA was 5% higher than that of the simple average.

One possible reason why EBMA might be more useful in the domain of weather forecasting is that weather forecasters can draw on massive datasets with little measurement error. In comparison, social scientists are often dealing with few observations and messy datasets, a situation that limits the performance of sophisticated statistical methods such as EBMA, which need to estimate parameters from a given sample of data.

The results reported in Table 1 suggest that one should be cautious when adopting complex methods that have been developed in distant fields and under different conditions. In any case, one should evaluate the performances of such methods relative to simple, robust, evidence-based alternatives. In the following section, we add such an evaluation to MHW's analysis of the performance of EBMA for forecasting US presidential elections.

4. A reassessment of EBMA for election forecasting

This section describes the data and the method used for calculating the forecasts, and provides the results of the analysis. All data and calculations are publicly available at the Harvard Dataverse (Graefe, 2014).

4.1. Data and method

The data and code were obtained from MHW so as to allow us to replicate their original findings for the task of US presidential election forecasting.[3] The dataset includes data for six established models (whose forecasts are published regularly in political science journals) up until the 2008 election. The six models were developed by Abramowitz (2008), Campbell (2008), Erikson and Wlezien (2008), Fair (2009), Hibbs (2000), and Lewis-Beck and Tien (2008). In order to gain one more observation for testing the out-of-sample accuracies, we also added the figures from the 2012 election.

All of the models – and thus also the combined forecasts – provide forecasts of the incumbent party's national two-party popular vote share.

4.1.1. EBMA forecasts

The original MHW code was used to compute the EBMA forecasts. MHW calculated the EBMA point forecasts as the median of the probability distribution.

[3] The MHW data are available at http://hdl.handle.net/1902.1/17286.

The EBMA predictions cannot be considered out-of-sample, because of an overlap between the training and validation sets that were used to estimate the component models and the EBMA parameters. The validation set covered all available observations for three models, and all but one observation for two models. That is, EBMA essentially used the in-sample accuracy of five out of six models to estimate the component weights.

An adequate accuracy test would require the component models to be trained (i.e., estimated) on the training set before the EBMA weights were estimated based on their out-of-sample accuracies in the validation set. Due to their limited dataset, MHW were unable to split the sample into separate training, validation, and test sets.

4.1.2. Benchmark forecast: simple averages

The simple unweighted average is the most widely accepted benchmark for combining forecasts (Armstrong, 2001; Clemen, 1989; Graefe et al., 2014). We therefore calculated the simple average across the individual forecasts from all six models in each election year in order to assess the relative accuracy of EBMA.

Combined forecasts based on the simple average were calculated as one-election-ahead predictions, and are therefore out-of-sample or pseudo ex ante. That is, only data that *would* have been available at the time of the particular election being forecast were used to estimate the model. For example, when producing a forecast for the 2004 election, only data up to the 2000 election were used to estimate the model. When producing a forecast of the 1984 election, only data up to 1980 were used, and so on.

4.1.3. Time horizon and error measure

The forecast accuracy was analyzed across the ten US presidential elections from 1976 to 2012. The absolute error was used to measure the absolute deviation of the point forecast from the actual election result.

4.2. Results

We first analyze the relationship between the individual models' past and future predictive accuracies. Then, we look at the predictive value of the weights estimated by EBMA for the individual models. Finally, we compare the relative accuracies of the combined forecasts based on EBMA and the simple average.

4.2.1. Predictive value of historical accuracy

As was noted in Section 2, past accuracy is often a poor indicator of a model's future accuracy, and the forecasting of US presidential elections is no exception. Holbrook (2010) reports that the accuracies of nine established econometric models varied considerably across the three elections from 1996 to 2004, with the models that were among the most accurate in one election often being among the least accurate in another.

We extended Holbrook's analysis to the six models in our dataset and the nine elections from 1980 to 2012. For each election, we calculated Spearman's rank correlation between the models' absolute errors in that year and their absolute errors in the previous election. In six of the nine elections, the correlation was negative, and the mean correlation across the nine elections was −0.20. These results confirm those of Holbrook (2010): the models that were among the most accurate in a given election tended to be among the least accurate in the succeeding election. For a more conservative estimate and to protect against outliers, we also calculated the correlation between the models' absolute errors in each of the five elections from 1996 to 2012 and their mean absolute errors across the previous five elections. Again, the correlation was negative in four of the five elections, with a mean correlation of −0.22. Thus, there appears to be a negative relationship between current and historical accuracy levels, or at best no relationship.

4.2.2. Validity of weights estimated by EBMA

The variance in accuracy across models and election years should make it difficult to estimate "optimal" weights for combining forecasts. In such situations, methods such as EBMA, which estimate weights using data on the components' historical accuracy levels, might even harm forecast accuracy.

The Appendix shows the weights that EBMA assigned to each of the individual models when calculating forecasts for each of the ten elections between 1976 and 2012. The model weights were highly unstable and varied considerably across elections, which is a sign of overfitting.[4] In general, there was no relationship between the weight that EBMA assigned to a model for forecasting a particular election and the model's corresponding absolute forecast error in that election ($r = -0.06$). In other words, the model weights estimated by EBMA were no indicator of the models' forecast accuracy.

We also looked at how many individual component forecasts entered the calculation of the combined EBMA forecasts. In no election did EBMA combine the forecasts from all six models. On average across the ten elections, EBMA used only 3.3 out of six model forecasts, and thus ignored a considerable amount of information.

4.2.3. Relative accuracies of simple averages and EBMA

Table 2 shows the mean absolute errors for both combining methods for each of the ten elections from 1976 to 2012 and overall. The simple average yielded a lower absolute error than EBMA in seven of the ten elections, while EBMA was more accurate in the three remaining elections. Across the ten elections, EBMA yielded a MAE of 1.7 percentage points, compared to a MAE of 1.3 percentage points for the simple average. That is, on average, the error of EBMA was 33% (or 0.4 percentage points) higher than the error of the simple average.

[4] For example, the Fair model was excluded from the combined EBMA forecast in six of the ten elections. However, the same model also received the largest weight of all models in two elections (1976 and 2004), and the second largest weight in one election (2000). Also, when taken together, the models by Fair and Hibbs account for more than half (56%) of the weight across the ten elections—an ironic result, given that, overall, these two models provided the least accurate forecasts of all models. Also, in

Table 2
Absolute errors of forecast combinations based on EBMA and the simple average, with error ratios.

Election year	EBMA	Simple average	Error ratio
1976	2.52	1.11	2.27
1980	0.39	0.27	1.47
1984	0.77	0.14	5.50
1988	0.97	0.90	1.08
1992	0.91	3.07	0.30
1996	3.02	0.54	5.59
2000	2.02	2.58	0.78
2004	2.09	2.24	0.93
2008	0.53	0.37	1.43
2012	4.01	1.69	2.37
Mean	**1.72**	**1.29**	**1.33**

5. Discussion

This analysis of US presidential election forecasting shows that EBMA failed to estimate robust model weights due to the large uncertainty about the relative accuracies of the individual models, which varied widely across the forecast horizon. As a result, EBMA provided point forecasts that were 33% (or 0.4 percentage points) less accurate, on average, than simply taking the average of all available forecasts. This finding is consistent with extensive prior research on combining, which has repeatedly found that simple averages yield predictions that are difficult for more complex approaches to beat.

5.1. Data limitations when predicting social science problems

The results of the present study probably *under*estimate the gains from using simple averages rather than EBMA for US presidential election forecasting, and should therefore be considered a low boundary. The reason for this is that the EBMA forecasts cannot be considered to be out-of-sample, as they used in-sample information about the relative accuracies of the models for the estimation of the component weights. In comparison, the combined forecasts based on simple averages are out-of-sample, as they used only data that would have been available at the time of a particular election.

The limited amount of data available for US presidential elections does not allow for the sample being split into separate training, validation, and test sets. Thus, US presidential election forecasting at the national level is unsuited for the application of EBMA. However, it can be difficult to choose the relative sizes of the three data sets even for large samples.[5] The EBMA literature does not provide clear guidelines for how to make such difficult decisions. MHW (p. 277) note that "the only general rule is: it depends", and Raftery et al. (2005) specify the need for an automatic procedure for choosing the training and validation sets. The lack of clear guidelines is problematic, as it introduces "researcher degrees of freedom"; that is, analysts can make decisions that are to some extent arbitrary, which may enable them to introduce their own biases.

The results demonstrate that one should exercise caution when adopting complex methods and applying them to social science problems, even if the methods have proved successful in other fields. EBMA might work well in data-heavy applications such as weather forecasting, where one can build on large and reliable datasets, but many problems in the social sciences are restricted by limited and noisy data. The present analysis suggests that EBMA is of limited value, and might even be harmful, under such conditions.

However, that being said, there might be situations in which EBMA's ability to assess the validity of forecasts automatically and to exclude useless forecasts will be superior to simply averaging all available forecasts. Further research is necessary to determine whether, and if so, *under what conditions*, EBMA is useful for social science problems. However, unless there is strong evidence supporting complex approaches such as EBMA, there is little need to depart from simple alternatives such as the widely accepted benchmark of using simple averages.

5.2. Barriers to using the simple average

Occam's Razor advises researchers to prefer simple models unless the lack of simplicity is offset by a greater explanatory power. Since Occam, many famous researchers have advocated the use of simple models. Albert Einstein is reputed to have said that "everything should be made as simple as possible but not simpler". Zellner (2004), who coined the phrase "keep it sophisticatedly simple", named several Nobel laureates as proponents of simplicity.

There is one problem, however: people have no faith in simple methods. Simple methods often face resistance, as people *wrongly* believe that complex solutions are necessary in order to solve complex problems. Hogarth (2012) reviews four influential studies which showed that simple methods often perform better than more complex ones. In each case, fellow researchers initially resisted the findings regarding the superiority of simple methods.

People are persuaded by complexity, possibly because they become impressed by what they do not understand. One experiment sent abstracts of two published papers (one in evolutionary anthropology and one in sociology) to 200 participants who had at least a postgraduate degree, and therefore experience in reading research papers. One or other of the abstracts was manipulated by adding a sentence from a completely unrelated paper. This sentence contained a mathematical equation that made no sense in the context. When participants with degrees in humanities, the social sciences or other related fields (e.g., education) were asked to judge the quality of the research,

four of the ten elections (1996, 2004, 2008, and 2012), the model that achieved the largest weight in the EBMA forecast was eventually the least accurate; and in another four elections (1980, 1992, 1996, and 2004), models that were excluded from the ensemble would have provided the most accurate individual forecasts. See the Appendix for details.

[5] Ulfelder (2013) was unable to find a reasonable way to divide up his dataset because of the high level of uncertainty about the underlying patterns over time (see also footnote 2).

they assigned a higher rating to the abstract that contained the meaningless mathematical equation. In comparison, this "nonsense math effect" was not observed for participants from fields with a stronger mathematical training (i.e., with degrees in mathematics, science, and technology), who were probably impressed less by the apparent complexity (Eriksson, 2012).

Awe and over-appreciation of complexity is also present in scientific publishing. Consequently, many researchers, especially young ones who are looking for tenure, think that they can increase their chances of being published by using complex methods. Sadly, papers that use complex methods might indeed be easier to publish than those using simple methods, as the fact that simple methods are easy to understand means that they are also easy to criticize. For example, combining forecasts by calculating simple averages is often disparaged as being naïve and atheoretical.

However, the use of simple averages is well grounded in statistical theory. In pre-specifying equal weights for all component forecasts, the analyst decides to ignore the components' relative accuracies and to deliberately introduce a bias that reduces variance. While a lower variance generally reduces a model's ability to *explain* given data, it avoids the danger of overfitting a model to the data. Thus, a low variance can be beneficial when predicting new data; for example, in situations that involve a considerable level of uncertainty. In statistical theory, this is known as the bias–variance tradeoff (Hastie et al., 2001).

In general, a model's predictive performance depends on the bias and variance components of the forecast error, which depend on the conditions of the forecasting problem (Gigerenzer & Brighton, 2009). Forecasters who face highly uncertain environments (e.g., due to ambiguity about causal relationships, external shocks, or the existence of noisy data) should acknowledge this uncertainty. That is, they should incorporate their prior knowledge (i.e., that predictions about the future are difficult to make in such situations) into their model. This provides the theoretical rationale for using the simple average when combining forecasts, as the method introduces conservatism. In fact, being conservative is one of the most effective ways to create accurate forecasts (Armstrong, Green, & Graefe, in press).

The combined EBMA forecasts for predicting US presidential elections analyzed in the present study were not conservative. First, the weights assigned to the component models were highly unstable and varied greatly across elections. Second, the combined EBMA forecasts ignored considerable amounts of information. In each election, at least one model was excluded from the combined EBMA forecasts. On average across the ten elections, EBMA ignored almost half of the available information. While a reliance on fewer models increases the variance in the forecast, the risk of overfitting – and thus the risk of a poor out-of-sample accuracy – increases too. In comparison, the simple average is conservative, as it acknowledges the uncertainty in the environment and uses all of the available information, which makes the method more robust when predicting new data.

6. Conclusions

Variants of Bayesian model averaging have become increasingly popular over the past few years. In December 2014, a Google Scholar search of ("Bayesian model averaging" AND forecasting) yielded more than 4900 hits, of which 2700 (55%) can be attributed to papers published in the past four years. However, we had difficulty in finding tests of predictive validity. While the method might work well in data-heavy domains such as weather forecasting, the evidence from economic and political forecasting suggests that EBMA is little more (or even less) accurate than calculating simple averages.

The present study does not intend to suggest that EBMA cannot be useful for forecasting social science problems. At the very least, EBMA is valuable in that it promotes the principle of combining forecasts. In most situations, people will make better predictions by combining forecasts, rather than by relying on a single source. However, forecasters need to assess whether EBMA is an appropriate choice given the conditions of the forecasting problem. Many social science problems are restricted by limited and noisy data, a situation that calls for robust methods that avoid overfitting.

Over the past three decades, forecasting research has achieved many advances, due to the strong emphasis on empirical comparisons of alternative methods (Armstrong et al., in press). Researchers who develop new methods (or adopt methods that were developed in a different field) need to compare the methods' out-of-sample performances to those of widely accepted benchmarks, and journals should require them to provide such analyses. Otherwise, the "bias towards complexity-for-the-sake-of-complexity (and tenure)" is likely to crowd out simpler and more robust methods (Schrodt, 2014, p. 292).

Acknowledgments

Scott Armstrong, Alfred Cuzán, Andrew Gelman, Thomas Gneiting, Randall Jones, Geoff Kenny, Carl Klarner, Adrian Raftery, Jay Ulfelder, Heiko von der Gracht, Christopher Wlezien, and Jonathan Wright provided helpful comments and suggestions. We also received valuable comments when presenting this manuscript at the *2014 International Symposium on Forecasting* in Rotterdam and the *2014 ECPR General Conference* in Glasgow. To ensure that the evidence has been summarized properly, and to confirm that no relevant evidence had been overlooked, we contacted the authors of all key articles for whom we found valid email addresses. Responses were received in all but three cases, which led to further improvements. Mario Haim replicated the results independently. Any errors remain the responsibility of the authors. This work was supported by an LMUexcellent research fellowship from the Center for Advanced Studies at LMU Munich.

Appendix. EBMA weights (in %) and absolute errors (AE) of the six individual models used in the calculation of the EBMA forecast

See the table in Box I.

		Hibbs	Campbell	Fair	Abramowitz	Erikson and Wlezien	Lewis-Beck and Tien
1976	Weight	27		50	23		
	AE	3.1	1.1	2.5	0.8	3.4	4.7
1980	Weight	56		13	31		#
	AE	0.8	2.6	0.9	1.8	0.6	0.4
1984	Weight	58	15		27		
	AE	1.1	0.6	4.0	0.5	1.7	2.7
1988	Weight	57	13		30		
	AE	1.5	0.1	2.6	1.9	0.7	1.8
1992	Weight	66	34				#
	AE	1.1	0.6	9.2	2.0	5.7	0.0
1996	Weight	**66**	32		2	#	
	AE	**3.4**	3.2	1.1	3.0	0.3	1.9
2000	Weight		33	32	3	16	15
	AE	6.3	2.9	1.6	3.9	2.1	1.8
2004	Weight	12	40	**48**			#
	AE	1.5	0.5	**4.8**	2.2	4.8	0.4
2008	Weight	25	**36**		6	17	17
	AE	1.4	**6.3**	2.0	2.4	0.1	2.6
2012	Weight	**50**	26			24	
	AE	**5.7**	0.4	2.8	0.8	1.6	3.5
Mean	Weight	42	23	14	12	6	3
	AE	2.6	1.8	3.2	1.9	2.1	2.0

Empty cells indicate forecasts that achieved a weight of zero and thus were excluded from the calculation of the EBMA forecast.

#: Forecasts that achieved zero weight and thus were excluded from the calculation of the EBMA forecast, but turned out to be most accurate.

Bold: Forecasts that achieved the highest weight but turned out to be least accurate.

Box I.

References

Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science and Politics*, *41*(4), 691–695.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners* (pp. 417–439). New York: Springer.

Armstrong, J. S., Green, K. C., & Graefe, A. (2014). Golden rule of forecasting: be conservative. *Journal of Business Research*, in press. Available at: goldenruleofforecasting.com.

Bork, L., & Møller, S. V. (2015). Forecasting house prices in the 50 states using dynamic model averaging and dynamic model selection. *International Journal of Forecasting*, *31*(1), 63–78.

Campbell, J. E. (2008). The trial-heat forecast of the 2008 presidential vote: performance and value considerations in an open-seat election. *PS: Political Science and Politics*, *41*(4), 697–701.

Clark, T. E., & McCracken, M. W. (2010). Averaging forecasts from VARs with uncertain instabilities. *Journal of Applied Econometrics*, *25*(1), 5–29.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Erikson, R. S., & Wlezien, C. (2008). Leading economic indicators, the polls, and the presidential vote. *PS: Political Science and Politics*, *41*(4), 703–707.

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making*, *7*(6), 746–749.

Fair, R. C. (2009). Presidential and congressional vote-share equations. *American Journal of Political Science*, *53*(1), 55–72.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: can anything beat the simple average? *International Journal of Forecasting*, *29*(1), 108–121.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Graefe, A. (2014). Replication data for: Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems. *Harvard Dataverse Network*: http://dx.doi.org/10.7910/DVN/EBMA.

Graefe, A. (2015). Improving forecasts using equally weighted predictors. *Journal of Business Research*, in press. Available at SSRN: http://ssrn.com/abstract=2311131.

Graefe, A., Armstrong, J. S., Jones, R. J., Jr., & Cuzán, A. G. (2014). Combining forecasts: an application to elections. *International Journal of Forecasting*, *30*(1), 43–54.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hibbs, D. A. (2000). Bread and peace voting in US presidential elections. *Public Choice*, *104*(1), 149–180.

Hogarth, R. M. (2012). When simple is hard to accept. In P. M. Todd, G. Gigerenzer, & A. R. Group (Eds.), *Ecological rationality: intelligence in the world* (pp. 61–79). Oxford: Oxford University Press.

Holbrook, T. M. (2010). Forecasting US presidential elections. In J. E. Leighley (Ed.), *The Oxford handbook of American elections and political behavior* (pp. 346–371). Oxford: Oxford University Press.

Lewis-Beck, M. S., & Tien, C. (2008). The job of president and the jobs model forecast: Obama for '08? *PS: Political Science and Politics*, *41*(4), 687–690.

Montgomery, J. M., Hollenbach, F., & Ward, M. D. (2012). Improving predictions using ensemble Bayesian model averaging. *Political Analysis*, *20*(3), 271–291.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174.

Schrodt, P. A. (2014). Seven deadly sins of contemporary quantitative political analysis. *Journal of Peace Research*, *51*(2), 287–300.

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

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430.

Stock, J. H., & Watson, M. W. (2005). *An empirical comparison of methods for forecasting using many predictors*. Working paper. Available at: www4.ncsu.edu/~arhall/beb_4.pdf.

Ulfelder, J. (2013). A multimodel ensemble for forecasting onsets of state-sponsored mass killing. *APSA 2013 annual meeting paper*. Available at: ssrn.com/abstract=2303048.

Wright, J. H. (2008). Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics*, *146*(2), 329–341.

Wright, J. H. (2009). Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting*, *28*(2), 131–144.

Zellner, A. (2004). Keep it sophisticatedly simple. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, inference and modelling: keeping it sophisticatedly simple* (pp. 242–262). Cambridge: Cambridge University Press.