Say Yes to the Guess:
Fitting Quality Ensembles on a Tight (data) Budget[*]

Jacob M. Montgomery
Department of Political Science
Washington University in St. Louis
Campus Box 1063, One Brookings Drive
St. Louis, MO, USA, 63130-4899

Florian M. Hollenbach
Department of Political Science
Duke University
Perkins Hall 326 Box 90204
Durham, NC, USA, 27707-4330

Michael D. Ward
Department of Political Science
Duke University
Perkins Hall 326 Box 90204
Durham, NC, USA, 27707-4330
corresponding author: michael.d.ward@duke.edu

August 13, 2012

[*] This work was partially supported by the Information Processing Technology Office of the Defense Advanced Research Projects Agency via a holding grant to the Lockheed Martin Corporation, Contract FA8650-07-C-7749. The current support is partially from the Office of Naval Research via ONR contract N00014-12-C-0066 to Lockheed Martin's Advanced Technology Laboratories.

Abstract

We consider ensemble Bayesian model averaging in the context of missing data. If ensembles members have few missing estimates the standard approaches in dealing with the missing observations, introduced by Fraley, Raftery and Gneiting (2010), work well . However, data in the social sciences generally do not full fill this requirement. Often missing predictions are neither random, nor rare. If component models have more extensive missing-ness, then EBMA has a tendency to overweight ensembles with a few observations, which can seriously undermine the advantages of using an ensemble approach in prediction. We demonstrate this problem and provide a solution that diminishes this possibility by introducing a "wisdom of the crowds" parameter. We demonstrate that this helps the predictive accuracy of EBMA estimates in political and economic applications in which there are ongoing forecasting efforts.

1 Introduction

Although accurate prediction of future events is not the primary goal for most social sciences, recent years have witnessed spreading of systematic forecasting from more traditional topics (e.g., GDP growth and unemployment) to many new domains (e.g., elections and mass killings) . Several factors have motivated this increase. To begin with, testing systematic predictions about future events against observed outcomes is generally seen as the most stringent validity check of statistical and theoretical models. In addition, forecasting of important political, economic, and social events is of great interest to policymakers and the general public who are generally less interested in testing theories of the world than correctly anticipating and altering the future.

With the proliferation of forecasting efforts, however, comes a need for sensible methods to aggregate and utilize the various scholarly efforts. One attractive solutions to this problem is to combine various prediction models and create an ensemble forecast. Combining forecasts reduces reliance on any single data source or methodology, but also allows for the incorporation of more information than any one model is likely to include in isolation. Across subject domains, ensemble predictions are usually more accurate than any individual component model. Second, they are significantly less likely to make dramatically incorrect predictions (Bates and Granger 1969; Armstrong 2001; Raftery et al. 2005).

The idea of ensemble learning itself has a long history in the machine learning and nonparametric statistics community. The most thorough treatment is found in Hastie, Tibshirani and Friedman

(2009). A wide range of statistical approaches including neural nets, bagging, random forests, additive regression trees, boosting, and more may be properly considered ensemble approaches.

One ensemble method advocated recently for forecasting is ensemble Bayesian model averaging (EBMA). This method was first proposed by Raftery et al. (2005) to improve weather forecasts. It has recently been suggested as a useful method for the social sciences by Montgomery, Hollenbach and Ward (2012). In essence, EBMA creates a finite mixture model that generates a weighted average of forecasts. EBMA mixture models seek to collate the good parts of existing forecasting models, while also avoiding over-fitting to past observations or over-estimating our certainty about the future. The hope is for greater accuracy as both the knowledge and implied uncertainty of a variety of approaches are integrated into a combined predictive probability distribution.

There are however several challenges for creating ensemble predictions for many applications in the social sciences. To begin with the amount and quality of data for calibrating ensembles is far from ideal. EBMA was first developed for use in weather forecasting where measurement of outcomes is fairly precise and data is relatively abundant. Predicting, for instance, water surface temperatures in 200 locations across five days provides 1,000 observations by which model weights can be calibrated. Forecasting quarterly GDP growth in the United States for five *years* only provides 20 observations.

A second and related issue is that in many forecasting exercises there tend to be a lot more forecasts than observations. For example, the well known forecaster of U.S. politics, Nate Silver, updates his forecasts of the 2012 presidential election weekly, yielding dozens of forecasts for a single outcome `http://fivethirtyeight.blogs.nytimes.com/`. Similarly, in the field of economics, a wide variety of consulting firms, banks, and international organizations each provide multiple forecasts for various economic quantities. One example are the various forecasts of the Federal Open Market Committee (FOMC) of the U.S. Federal Reserve Board, which are frequently updated, as for example here: `http://1.usa.gov/zjyisV`.

A final issue is the inconsistency with which forecasts are issued. Given the lengthy time periods involved, of any given time window there are many missing forecasts, especially in the social sciences. Moreover, we cannot assume that forecasts for any time period from a specific model or team are missing at random. Particularly unsuccessful forecasts may be suppressed or some forecasting efforts are over shorter time-periods than others. Moreover, forecasts have tended to accumulate with more observations being available for more proximate time periods.

One example of forecasting that combines all of these issues in the prediction of U.S. Presidential elections. Table 1 represents nearly an entirety of scholarly forecasts which produced more than one out-of-sample forecast for elections in the 20th century [1]. In this instance we have only five observations by which to calibrate an ensemble model while we have nine different forecasts models. Moreover, several of the individual forecasts are missing for a significant portion of the data, as their forecasting efforts started at a later date. The forecast of Cuzàn, for instance, is missing for 60% of the elections in this dataset.[2]

Table 1: Pre-election forecasts of the percent of the two-party vote going to the incumbent party in U.S. Presidential elections

	F	A	C	H	LBRT	L	Hol	EW	Cuz
1992	55.7	46.3	49.7	48.9	47.3				
1996	49.5	57.0	55.5	53.5	53.3		57.2	55.6	
2000	50.8	53.2	52.8	54.8	55.4	60.3	60.3	55.2	
2004	57.5	53.7	52.8	53.2	49.9	57.6	55.8	52.9	51.1
2008	48.1	45.7	52.7	48.5	43.4	41.8	44.3	47.8	48.1

Forecasts were published prior to each election by **F**air, **A**bramowitz, **C**ampbell, **H**ibbs, **L**ewis-**B**eck and **R**ice (1992), Lewis-Beck and **T**ien (1996-2008), **L**ockerbie, **Hol**brook, **E**rikson and **W**lezien and **Cuz**àn and Bundrick. Data were taken from the collation presented at `http://fivethirtyeight.blogs.nytimes.com/2012/03/26/models-based-on-fundamentals-have-failed-at-predicting-presidential-elections/`.

While particularly egregious for presidential forecasting as presented here, these data issues of missing observations and sparse data are endemic across the social sciences.

[1] (Fair 2009, 2011; Abramowitz 2008; Campbell 2008; Cuzàn and Bundrick 2004, 2008; Hibbs 2012; Lockerbie 2008; Erikson and Wlezien 2008; Graefe et al. 2010; Holbrook 2008, See, for example). A recent symposium in *PS: Political Science & Politics* presents and summarizes attempts by a variety of scholars to predict the 2012 U.S. presidential election. In the symposium contribution we have used the in-sample fitted values of the election forecasting efforts to calibrate the EBMA model. However, the strength of EBMA is greatest when the model is calibrated on true out-of-sample forecasts, thus we focus on these here.

[2] The predictions by Cuzàn for 2004 stems from the FISCAL model published prior to the 2004 election by Cuzàn and Bundrick (2004), while the 2008 prediction comes from the FPRIME short model presented in advance of the election (Cuzàn and Bundrick 2008). However, both models are quite similar in their composition.

In this paper, we explore several adjustments to the basic EBMA model as specified in Montgomery, Hollenbach and Ward (2012) that can help applied researchers create ensemble forecasts even in the presence of these kinds of data-quality issues. Specifically, we show EBMA can be adjusted to easily accommodate missing forecasts as first proposed by (Fraley, Raftery and Gneiting 2010). In addition, we propose an alteration to the basic model that can aid in the forecasting effort when the number of calibration observations are small and some component models suffer from lots of missing predictions. Below, we first briefly introduce the basic EBMA model in section 2. We then outline modifications to the model for missing-ness and small samples in sections 3 and 4. In section 5, we apply the adjusted EBMA model to unemployment data as well as presidential forecasting models shown in Table 1.

2 Notation and basic EBMA model

In this section we shortly summarize the basic notation and methods used to estimate ensemble Bayesian averaging models. Reader who are interested in the complete mathematics behind the method should consult Montgomery, Hollenbach and Ward (2012) and Raftery et al. (2005).

Assume a quantity of interest to forecast, $\mathbf{y}^{t^*}$, in some future period $t^* \in T^*$. Further assume that we have extant forecasts for events $\mathbf{y}^t$ for some past period $t \in T$ that were generated from K forecasting models or teams, $M_1, M_2, \ldots, M_K$, for which a prior probability distribution $M_k \sim \pi(M_k)$ exists. The PDF for $\mathbf{y}^t$ is denoted $p(\mathbf{y}^t|M_k)$. Under this model, the predictive PDF for the quantity of interest is $p(\mathbf{y}^{t^*}|M_k)$, the conditional probability for each model is $p(M_k|\mathbf{y}^t) = p(\mathbf{y}^t|M_k)\pi(M_k)/\sum\limits_{k=1}^{K} p(\mathbf{y}^t|M_k)\pi(M_k)$ and the and the marginal predictive PDF is $p(\mathbf{y}^{t^*}) = \sum\limits_{k=1}^{K} p(\mathbf{y}^{t^*}|M_k)p(M_k|\mathbf{y}^t)$. This can be viewed as the weighted average of the component PDFs where the weights are determined by each model's performance within the already-observed calibration period T.

2.1 Dynamic ensemble forecasting

The EBMA procedure assumes K forecasting models throughout the training (T') calibration (T) and test (T^*) periods that can be used as component models. These are calibrated in the training period T'. Optimally then the component model predictions for the calibration period T are made out-of-sample. The goal is to estimate the parameters for the ensemble prediction model using $\mathbf{f}_k^t$ for some period T. It is then possible to generate true ensemble forecasts ($\mathbf{f}_k^{t^*}$) for observations in the test period $t^* \in T^*$.

Let $g_k(\mathbf{y}|\mathbf{f}_k^{s|t,t^*})$ represent the predictive PDF of component k, which may be the original prediction from the forecast model or the bias-corrected forecast. The EBMA PDF is then a finite mixture of the K component PDFs, denoted $p(\mathbf{y}|\mathbf{f}_1^{s|t}, \ldots, \mathbf{f}_K^{s|t}) = \sum\limits_{k=1}^{K} w_k g_k(\mathbf{y}|\mathbf{f}_k^{s|t})$, where $w_k \in [0, 1]$ are model probabilities, $p(M_k|\mathbf{y}^t)$, and $\sum_{k=1}^{K} w_k = 1$. The ensemble predictive PDF with this notation is is then $p(y|f_1^{t^*}, \ldots, f_K^{t^*}) = \sum\limits_{k=1}^{K} w_k g_k(y|f_k^{t^*})$.

Past applications have statistically post-processed the predictions for out-of-sample bias reduction and treated these adjusted predictions as a component model. Raftery et al. (2005) propose approximating the conditional PDF as a normal distribution centered at a linear transformation of the individual forecast, $g_k(\mathbf{y}|\mathbf{f}_k^{s|t}) = N(a_{k0} + a_{k1}\mathbf{f}_k^t, \sigma^2)$. However, in the presence of sparse data, including the additional $\mathbf{a}$ parameters risks over-fitting and reduced predictive performance. We therefore use a simpler formulation where $g_k(\mathbf{y}|\mathbf{f}_k^t) = N(\mathbf{f}_k^t, \sigma^2)$. Thus, the ultimate predictive distribution for some observation y^{t^*} is

$$p(y|f_1^{s|t^*}, \ldots, f_K^{s|t^*}) = \sum_{k=1}^{K} w_k N(f_k^{t^*}, \sigma^2). \tag{1}$$

This, is a mixture of K normal distributions each of whose mean is determined by $f_k^{t^*}$ and which is scaled by the model weights w_k.

2.2 Parameter estimation

Since the component model forecasts, $f_1^t, \ldots, f_k^t$, are pre-determined, the EBMA model is fully specified by estimating model weights, $w_1, \ldots, w_k$ and the common variance parameter σ^2. We estimate these using maximum likelihood methods (Raftery et al. 2005), although Vrugt, Diks and Clark (2008) have proposed estimation via Markov chain Monte Carlo metods. The log likelihood function is

$$\mathcal{L}(w_1, \ldots, w_k, \sigma^2) = \sum_t log \left(\sum_{k=1}^{K} w_k N(f_k^t, \sigma^2) \right). \tag{2}$$

The log-likelihood function cannot be maximized analytically, so Raftery et al. (2005) propose an EM algorithm which explicitly expresses EBMA as a finite mixture model McLachlan and Peel (2000); Imai and Tingley (2012). We introduce the unobserved quantities z_k^t, which represents the probability that observation y^t is "best" predicted by model k. The E step involves calculating estimates for these unobserved quantities using the formula

$$\hat{z}_k^{(j+1)t} = \frac{\hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}{\sum\limits_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}, \tag{3}$$

where the superscript j refers to the jth iteration of the EM algorithm.

$w_k^{(j)}$ is the estimate of w_k in the jth iteration and $p^{(j)}(.)$ is shown in (1). Assuming these estimates of $z_k^{s|t}$ are correct, it is then straightforward to derive the maximizing value for the model weights. Thus, the M step estimates these as

$$\hat{w}_k^{(j+1)} = \frac{1}{n} \sum_t \hat{z}_k^{(j+1)t}, \tag{4}$$

where n represents the number of observations in the calibration dataset. Finally,

$$\hat{\sigma}^{2(j+1)} = \frac{1}{n} \sum_t \sum_{k=1}^{K} \hat{z}_k^{(j+1)t} (y - f_k^t)^2. \tag{5}$$

The E and M steps are iterated until the improvement in the log-likelihood is no larger than some pre-defined tolerance. We initiate the algorithm with the assumption that all models are equally likely, $w_k = \frac{1}{K} \, \forall \, k \in [1, \ldots, K]$ and $\sigma^2 = 1$.

3 Missing forecasts

The above described method however requires all component models to contain predictions for all observations in the calibration period. Thus if one model's observation at time t is missing, the only solution, given the algorithm above, is to list wise delete the observation in time t for all component models. To accommodate missing values in component models prediction within the EBMA procedure we follow Fraley, Raftery and Gneiting (2010) and modify the EM algorithm as follows.[3] Define

$$\mathcal{A}^t = \{i|\text{ensemble member i available at time t}\}.$$

.

which is simply the indicators of the list of components that provide forecasts for observation y_t. For convenience, define $\tilde{z}_k^{(j+1)t} \equiv \sum\limits_{k \in A^t} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t) / \sum\limits_{k \in A^t} w_k^{(j)}$. Equation 3 above is then replaced with

$$\hat{z}_k^{(j+1)t} = \begin{cases} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t)/\tilde{z}_k^{(j+1)t} \text{ if } k \in \mathcal{A}^t \\ 0 \text{ if } k \notin \mathcal{A}^t \end{cases} \tag{6}$$

The M steps in Equations 4 and 5 are likewise replaced with

$$\hat{w}_k^{(j+1)} = \frac{\sum\limits_t \hat{z}_k^{(j+1)t}}{\sum\limits_t \sum\limits_{k=1}^K \hat{z}_k^{(j+1)t}} \tag{7}$$

and

[3] In future research, we intend to compare alternative methods for handling missing data, including the use of gaussian copulas to impute the missing predictions (Hoff 2007).

$$\hat{\sigma}^{2(j+1)} = \frac{\sum_t \sum_{k=1}^{K} \hat{z}_k^{(j+1)t}(y - f_k^t)^2}{\sum_t \sum_{k=1}^{K} \hat{z}_k^{(j+1)t}}. \tag{8}$$

Thus in essence the likelihood is renormalized given the missing ensemble observations prior to maximization. Thus using the adjustments above, the EBMA algorithm now allows for missing observations in the component predictions.

4 Small sample adjustment

The second issue that is often problematic for combining forecasts in the social sciences in a principled way, is the small number of observations. When ensembles are calibrated on very few observations, there is an increased chance that EBMA may over-weight high performing models in a way that reduces out of sample performance. This is especially true when the short calibration period is combined with missing observations in the component model predictions.

In an attempt to deal with this issue, we introduce a "wisdom of crowds" parameter, $c \in [0, 1]$, that reflects our prior belief that all models should receive some weight. In essence, we rescale z_k^t to have a minimum value $\frac{c}{K}$. This essentially states that there is, at a minimum, a $\frac{c}{K}$ probability that the observation is correctly represented by each model k. Since $\sum_{k=1}^{K} z_k^t = 1$, this implies that $z_k^t \in [\frac{c}{K}, (1 - c)]$. To achieve this, we replace Equation 4 above with

$$\hat{z}_k^{(j+1)t} = \frac{c}{K} + (1 - c)\frac{\hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}{\sum_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t)}. \tag{9}$$

Note that when $c = 1$, that all models are considered equally informative about the outcome and $w_k = \frac{1}{K} \forall K$. Thus, we see that the arithmetic mean or median of component forecasts for time period t represents a special case of EBMA where $c = 1$.[4] Likewise, the general EBMA discussed

[4] The mean or median would be equivalent depending on if the posterior mean or median is used to make a point prediction.

in Montgomery, Hollenbach and Ward (2012) represents a special case of this more general model where $c = 0$.

5 Applications

After explaining the methodological background to EBMA in general and the two adjustments above, we now turn to examining how these methods work in two areas that typify forecasting in the social sciences. One is the estimation of an economic series, unemployment, and the second in the area of predicting the vote for the incumbent in U.S. presidential elections.

5.1 Quarterly unemployment

Forecasting macroeconomic variables is a quite common exercise in the field of economics and statistics. Receiving as accurate as possible forecasts of economic variables is a necessity for many policy makers as well as businesses. Most forecasts are created using a wide variety of statistical models.[5] The majority of scholars employs sophisticated time-series models in an attempt to make the most accurate predictions. For a long time the most commonly used statistical method for economic forecasts were relatively simple ARIMA and vector autoregressive (VAR) models. These models were designed in an attempt to deal with the inherent time-series dynamics in the data. However, the sophistication and complexity of forecasting models has increased considerably since the 1980s. In particular non-linear dynamic models have gained prominence, such as for example, threshold autoregressive (TAR), Markov switching autoregressive (MSA), or smooth transition autoregression (STAR) models (Elliott and Timmermann 2008; Montgomery et al. 1998). More recently Bayesian VAR and state-space models have also gained more attention of forecasters when predicting unemployment and other economic variables (De Goijer and Hyndman 2006; Elliott and Timmermann 2008).

In addition, beginning with Bates and Granger (1969) scholars have attempted to improve the

[5] For a more comprehensive overview on foresting of economic variables and time-series forecasting see Elliott and Timmermann (2008) and De Goijer and Hyndman (2006).

accuracy of forecasts by combining different predictions in a meaningful way (Palm and Zellner 1992; Elliott and Timmermann 2008). With the introduction of Bayesian averaging methods, the principled combination of predictive models has been successfully used to improve the efforts of forecasting inflation (Koop and Korobilis 2009; Wright 2009), GDP (Billio et al. 2010), stock prices (Billio et al. 2011) as well as exchange rates (Wright 2008).

In addition to statistical models, economic variables are often predicted using expert surveys. This is the case for the *Survey of Professional Forecasters (SPF)* published by the *Federal Reserve Bank of Philadelphia*, which published forecasts for a large number of economic variables in the US, including, but not limited to unemployment rate, inflation, and GDP. The SPF was first administered in 1968 by the American Statistical Association and the National Bureau of Economic Research (NBER). However, since 1990 it is conducted by the Federal Reserve Bank of Philadelphia.[6] Every first month of the quarter a survey is send out to the forecasters, which has to be returned by the middle of the second month of the quarter with the forecasters prediction. Forecasts are made for the current quarter as well as several quarters into the future.

This plethora of quarterly predictions is optimal for us to apply Ensemble Bayesian model averaging on. We thus use forecasts of the civilian unemployment rate (UNEMP) as published by the SPF. as component models. For this application we select the forecast horizon to be four quarters into the future, i.e. predictions made in the first quarter of 2002 are for the first quarter of 2003 and so on. In total the SPF data on unemployment contains forecasts by 569 different teams, however for any quarter in the SPF sample, the average number of forecast teams making a prediction for four quarters into the future is quite small and the majority of observations for any given quarter is missing.[7]

In addition to the forecasts collected in the survey, we include the “Greenbook” forecasts produced by the Federal Reserve. These forecasts are made by the research staff of the Board of Governors and are handed out prior to meetings of the Federal Reserve Open Market Committee

[6] See `http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/` for more information.

[7] On average only7.5 per cent of all teams make a forecast for any one quarter.

(FOMC). We merge these data based on the quarter that is predicted. However, it has been noted that an aggregation of the SPF forecasts and the Greenbook predictions are very similar to each other (Baghestani 2008).

To evaluate forecasts we use the most recent data available, i.e. the most recent vintage, where the data is likely to have been revised. All predictions are evaluated using the historical unemployment rate for each quarter as recorded today.[8]

Given the SPF and Greenbook unemployment forecasts we calibrate an ensemble model for each period t, using forecaster performance over the past ten quarters. Thus the EBMA calibration period is a rolling window, with a different calibration period for each quarter EBMA model. Only forecasts that had made predictions for five of the past ten quarters were included in the ensemble calibration. Thus, the EBMA model uses only 163 models out of a possible 293 forecasting models that made predictions during the period we study. Due to missing data early in the time series and the fact that Greenbook forecasts are sequestered for five years, we generate forecasts beginning in the third quarter of 1983 and running through the fourth quarter of 2007.

Figure 1 provides a visual representation of EBMA model calibrations throughout this period. As one can see, the EBMA model calibration is different for each quarter, given the rolling calibration window discussed above. The models depicted in this figure are estimated with the wisdom of crowds tuning parameter set to a modest $c = 0.05$. The colors indicate the model weight assigned to each component on a red-blue color ramp (components not included in the ensemble are simply blank). Component forecasts that did not receive any weights in a given model are shown in dark blue while models that are heavily weighted are shown in red.

Figure 1 shows clearly the difficulties inherent in forecasting with this type of data. As mentioned above, for any given year, only a subset of forecasting teams even offer a prediction. Further, an even smaller subset of models offer both a prediction for the quarter to be forecasted and have

[8] As Croushore and Stark (2001) describe, depending on the forecast exercise it can make a difference whether the forecast models are evaluated using "real-time" or the latest available data. We have decided here to use the latest available data and do not believe that it should make a difference in our case, as all predictions are evaluated against the same data and EBMA is a mixture of the other forecast models. However, in a future version of this paper we will replicate the results in this analysis using real-time data to evaluate the forecasts.

made a sufficiently large number of prior forecasts to facilitate the model calibration. Finally, the very sparseness of the data encourages the ensemble model to place a very large amount of weight on the best performing models.

We now turn to evaluating the performance of the ensemble relative to its 163 component forecasts. To do this, we focus on eight model fit indices available in the literature. The eight metrics we use are mean absolute error (MAE), root mean squared error (RMSE), median absolute deviation (MAD), root mean squared logarithmic error (RMSLE), mean absolute percentage error (MAPE), median absolute percentage error (MEAPE), median relative absolute error (MRAE) and percent worse (PW). The latter two metrics are measured relative to a naive model, simply predicting the future rate of unemployment as being the same as the current rate of unemployment.

It is important to note that many of these forecasters make predictions in a relatively small subset of cases. That is, each model k offers forecasts for only a subset of cases $n_k \subset n$. To create a fair comparison, therefore, we calculate these fit indices only for $n_k \forall k \in [1, K]$. By this measure, the EBMA model performs very well. Figure 2 provides a summary of these results. The top panel shows the percentage of metrics by which EBMA outperforms each component. The bottom panel shows the percentage of component models that EBMA "beats" as measured by each metric.

Notably, the relative superiority of EBMA to its components is somewhat less for components that provide few forecasts. This reflects the fact that with so many forecasts, some are likely to be more accurate than the ensemble by chance alone. Additionally, when the number of forecasts is low it is likely that a given model received less weight than it "deserves" given the model's performance. However, across a large number of forecasts, EBMA significantly outperforms any of its components, including the Greenbook (GB). It is also worth noting that only 6 out of the total 163 components outperform EBMA on every metric.

Another approach to evaluating the performance of EBMA is to compare its predictive accuracy to that made by other systematic forecasting efforts and methods of generating ensemble predictions. Specifically, we compare EBMA's predictive accuracy to (1) the Greenbook, (2) the median forecaster prediction and (3) the mean forecaster prediction.[9] The first three of these forecasts and the true level of unemployment are shown in Figure 3.

Table 2: Model Comparison via Error Statistics

	MAE	RMSE	MAD	RMSLE	MAPE	MEAPE	MRAE	PW
EBMA (c=0)	0.54	0.74	0.37	0.009	8.37	6.49	**0.73**	**27.36**
EBMA (c=0.05)	**0.54**	0.74	**0.37**	**0.009**	**8.33**	**6.30**	0.75	**27.36**
EBMA (c=0.1)	0.54	0.74	0.35	0.009	8.40	6.44	0.76	28.30
EBMA (c=1)	0.61	0.80	0.46	0.010	9.72	8.92	0.95	46.23
Greenbook	0.57	**0.73**	0.43	0.009	9.37	8.81	1.00	45.28
Forecast Median	0.62	0.81	0.47	0.011	9.83	8.87	0.98	47.17
Forecast Mean	0.61	0.80	0.46	0.010	9.71	9.06	0.93	46.23

This table depicts a number of error statistics for a variety of forecasting models calculated based on the SPF and Greenbook data. The model with the lowest score for each metric are shown in bold. As one can see aside from the RMSE and the MRAE statistics, the EBMA model with $c = 0.05$ scores the best on all other statistics.

Table 2 compares these baseline models using all eight of the metrics to EBMA models with c =0, 0.05, 0.1, and 1 respectively. The bolded cells in each column indicate the model that performed "best" as measured by each metric. With one exception, the Greenbook outperforms the ensemble by 0.01 on RMSE, the EBMA model outperforms both the Greenbook forecast and the unweighted mean and median forecast. Moreover, these results indicate that the c parameter is best set to a small number. In general, the model with $c = 0.05$ performs best (or is tied for best) on six out of eight of these metrics.

Figure 3 shows a visual representation of the Greenbook, median SPF and the EBMA (with $c = 0.05$) forecasts over time, as well as the true unemployment rate. As was noted above and is clearly visible, the SPF and Greenbook forecasts are quite similar. Baghestani (2008) noted that the Greenbook forecast is slightly biased to over predict the unemployment rate. In some periods EBMA is able to correct this bias, however given the similarity of component models, the improvement in that direction is rather small. In general however it is easily visible that the EBMA

[9] Note that the EBMA model is calculated on only a the subset of forecasts that have made a sufficiently large number of recent predictions to calibrate model weights. Thus, the median forecast and the ensemble forecast will not be the same even when $c = 1$.

forecast is closer to the actual rate than the median SPF or the Greenbook forecast.

5.2 U.S. presidential elections

Informed by the above discussion, we now turn to our second application and return briefly to the example with which we began – predicting U.S. presidential elections. As was explained above, the number of observations with which to calibrate the EBMA model is extremely small in this example, while quite a number of forecasting models exist. This created problems for the original EBMA algorithm and warrants our proposed adjustments.

Using the forecasts shown in Table 1, we fit an EBMA model with $c = 0.05$. The model weights and calibration fit statistics for the ensemble and its components are shown in Table 3.

Table 3: Presidential Election Forecast: Model weights and calibration period fit statistics.

	W	rmse	mae
EBMA		1.92	1.56
F	0.02	5.53	4.58
A	0.78	2.02	1.72
C	0.07	3.46	2.88
H	0.04	2.68	2.44
LBRT	0.06	2.78	2.28
L	0.00	7.33	6.97
Hol	0.01	5.73	4.77
EW	0.02	2.74	2.25
Cuz	0.00	1.27	0.95

As can be seen in table 3, the EBMA model assigns the majority of weight to the Abramowitz model with the model by Campbell receiving the second largest weight. These weights are based on the performance of each model in forecasting the incumbent vote share in the presidential elections between 1992 and 2008. The Cuzàn and Bundrick model is weighted to such a small degree because only out-of-sample predictions for 2004 and 2008 were available here.

Figure 4 shows the posterior predictive distribution for the 2008 election (top) and, based on current forecasts from each of the component models, the out-of-sample prediction for the 2012 election. We predict that Obama is going to win by very little, however the credible intervals are quite wide, indicating a lot of uncertainty. Component models predictive distributions are shown in color (scaled by their respective weight), while the EBMA predictive distribution is shown in black. Vertical dashes indicate the point prediction of each model (bold dash for the EBMA model). The vertical dashed line in the top panel depicts the actual election result in 2008.

6 Discussion

Ensemble Bayesian model averaging is a principled way of combining forecasts to improve prediction accuracy. However, the calibration of such models in the social sciences is often hindered by the quality as well as availability of data. For one, in many forecasting exercises the number of forecasting models is large, yet the number of observations on which the EBMA model can be trained is small. This creates problems for the estimation of model weights, as it is likely that overly high weights are assigned to models that are performing well over this particular period. This is especially true should the EBMA model be calibrated on in-sample forecasts of the component models. Second, many predictive models do not provide forecasts for all observations in the sample, as some forecasts may be missing or the time-period for which forecasts were made are different for different models. In the standard EBMA model introduced in Montgomery, Hollenbach and Ward (2012) missing observations in component model predictions necessitate list-wise deletion.

In this paper we attempt to deal with both of these issues, to make EBMA more applicable for researchers and predictioneers in the social sciences. After introducing the math behind the common EBMA framework and its notation in the first section, we proceed to introduce an adjustment to the EBMA estimation that allows for missing observations in the calibration period. As a further adjustment we then propose to introduce a “wisdom of the crowds” parameter into the model, which forces EBMA to put some weight on all component models. Adding this constant aids the calibration of EBMA when the number of observations in the calibration period is small.

After explaining our adjustments we apply the "new" EBMA model to two prediction exercises. In section 5.1 we use ensemble Bayesian model averaging to combine predictions of the unemployment rate in the US from the Survey of Professional Forecasters as well as the Fed's Greenbook. As we show, even when a large number of forecasts is missing for any given quarter, EBMA generally outperforms the Greenbook, SPF component models, as well as the median and mean SPF forecast. In a second example we use the out-of-sample forecasts of nine famous prediction models of presidential elections from 1992 to 2008 to calibrate an ensemble model. We use the model calibrated to make an informed prediction for the 2012 elections based on a weighted combination of the component model predictions for 2012. According to the EBMA model we expect President Obama to win the popular vote in the 2012 election by a small margin, however the uncertainty estimated around the prediction is quite large.

Missing data is always a conundrum, and a pain. This is especially true when creating ensemble predictions. The combination of missing observations with short calibration periods is especially damaging. EBMA typically underweights the ensembles with a lot of missing observations, and as a result can diminish, rather than enhance, the predictive accuracy of the weighted average. We introduce a way around this problem by the introduction of a parameter which spreads the weights out over the ensemble components in a way that helps to preserve the advantages of the ensemble.

In future drafts of this paper we hope to (1) compare alternative methods of handling missing data (2) discuss how to select a window of time for calibration and (c) conduct some simulation studies to explore settings for c parameter and to test the numerical stability of our results.

Thank you and good night. Tip your server.

Appendix

Mathematical description of the various model fit statistics here.

References

Abramowitz, Alan I. 2008. "Forecasting the 2008 Presidential Election with the Time-for-Change Model." *PS: Political Science & Politics* 41(4):691–695.

Armstrong, J. Scott. 2001. Combining Forecasts. In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, ed. J. Scott Armstrong. Norwell, MA: Kluwer Academic Publishers.

Baghestani, Hamid. 2008. "Federal Reserve versus private information: Who is the best unemployment rate predictor?" *Journal of Policy Modeling* 30(1):101–110.

Bates, J.M. and Clive W.J. Granger. 1969. "The Combination of Forecasts." *Operations Research* 20(4):451–468.

Billio, Monica, Roberto Casarin, Francesco Ravazzolo and Herman K. Van Dijk. 2010. "Combining Predictive Densities Using Bayesian Filtering with Applications to US Economics Data." Norges Bank Working Paper. `http://ssrn.com/abstract=1735421` (accessed June 1, 2011).

Billio, Monica, Roberto Casarin, Francesco Ravazzolo and Herman K. Van Dijk. 2011. "Bayesian Combinations of Stock Price Predictions with an Application to the Amsterdam Exchange Index." Tinbergen Institute Discussion Paper No. 2011-082/4. `http://www.tinbergen.nl/discussionpapers/11082.pdf` (accessed June 1, 2011).

Campbell, James E. 2008. "The Trial-heat Forecast of the 2008 Presidential Vote: Performance and Value Considerations in an Open-Seat Election." *PS: Political Science & Politics* 41(4):697–701.

Croushore, Dean and Tom Stark. 2001. "A Real-Time Data Set for Macroeconomists." *Journal of Econometrics* 105(1):111–130.

Cuzàn, Alfred G. and Charles M. Bundrick. 2004. "Fiscal Effects on Presidential Elections: A Forecast for 2004." Paper prepared for presentation at the American Political Science Association, Chicago `http://uwf.edu/govt/facultyforums/documents/fiscaleffectsprselect2004.pdf`.

Cuzàn, Alfred G. and Charles M. Bundrick. 2008. "Forecasting the 2008 Presidential Election: A Challenge for the Fiscal Model." *PS: Political Science & Politics* 41(4):717–722.

De Goijer, Jan G. and Rob J. Hyndman. 2006. "25 years of time series forecasting." *International Journal of Forecasting* 22(3):443–473.

Elliott, Graham and Allan Timmermann. 2008. "Economic Forecasting." *Journal of Economic Literature* 46(1):3–56.

Erikson, Robert S. and Christopher Wlezien. 2008. "Leading Economic Indicators, the Polls, and the Presidential Vote." *PS: Political Science & Politics* 41(4):703–707.

Fair, Ray C. 2009. "Presidential and Congressional Vote-Share Equations." *American Journal of Political Science* 53(1):55–72.

Fair, Ray C. 2011. "Vote-Share Equations: November 2010 Update." Working Paper, Yale University. http://fairmodel.econ.yale.edu/vote2012/index2.htm (accessed March 07, 2011).

Fraley, Chris, Adrian E. Raftery and Tilmann Gneiting. 2010. "Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging." *Monthly Weather Review* 138(1):190–202.

Graefe, Andreas, Aldfred G. Cuzan, Randal J. Jones and J. Scott Armstrong. 2010. "Combining Forecasts for U.S. Presidential Elections: The PollyVote." Working Paper. http://dl.dropbox.com/u/3662406/Articles/Graefe_et_al_Combining.pdf (accessed May 15, 2011).

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hibbs, Douglas A. 2012. "Obama's Re-election Prospects under 'Bread and Peace/ Voting in the 2012 US Presidential Election." http://www.douglas-hibbs.com/HibbsArticles/HIBBS_OBAMA-REELECT-31July2012r1.pdf.

Hoff, Peter D. 2007. "Extending the Rank LIkelihood for Semiparametric Copula Estimation." *Annals of Applied Statistics* 1(1):265–283.

Holbrook, Thomas M. 2008. "Incumbency, National Conditions, and the 2008 Presidential Election." *PS: Political Science & Politics* 41(4):709–712.

Imai, Kosuke and Dustin Tingley. 2012. "A Statistical Method for Empirical Testing of Competing Theories." *American Journal of Political Science* 56(1):218–236.

Koop, Gary and Dimitris Korobilis. 2009. "Forecasting Inflation Using Dynamic Model Averaging." Working Paper. http://personal.strath.ac.uk/gary.koop/koop_korobilis_forecasting_inflation_using_DMA.pdf (accessed May 25, 2011).

Lockerbie, Brad. 2008. "Election Forecasting: The Future of the Presidency and the House." *PS: Political Science & Politics* 41(4):713–716.

McLachlan, Geoffrey and David Peel. 2000. *Finite Mixture Models*. New York, NY: John Wiley & Sons, Ltd.

Montgomery, Alan L., Victor Zarnowitz, Ruey Tsay and Tiaom George. 1998. "Forecasting the U.S. Unemployement Rate." *Journal of the American Statistical Association* 93(442):478–493.

Montgomery, Jacob M., Florian Hollenbach and Michael D. Ward. 2012. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20(3):271–291.

Palm, Franz C. and Arnold Zellner. 1992. "To Combine of Not to Combine? Issues of Combining Forecasts." *Journal of Forecasting* 11(8):687–701.

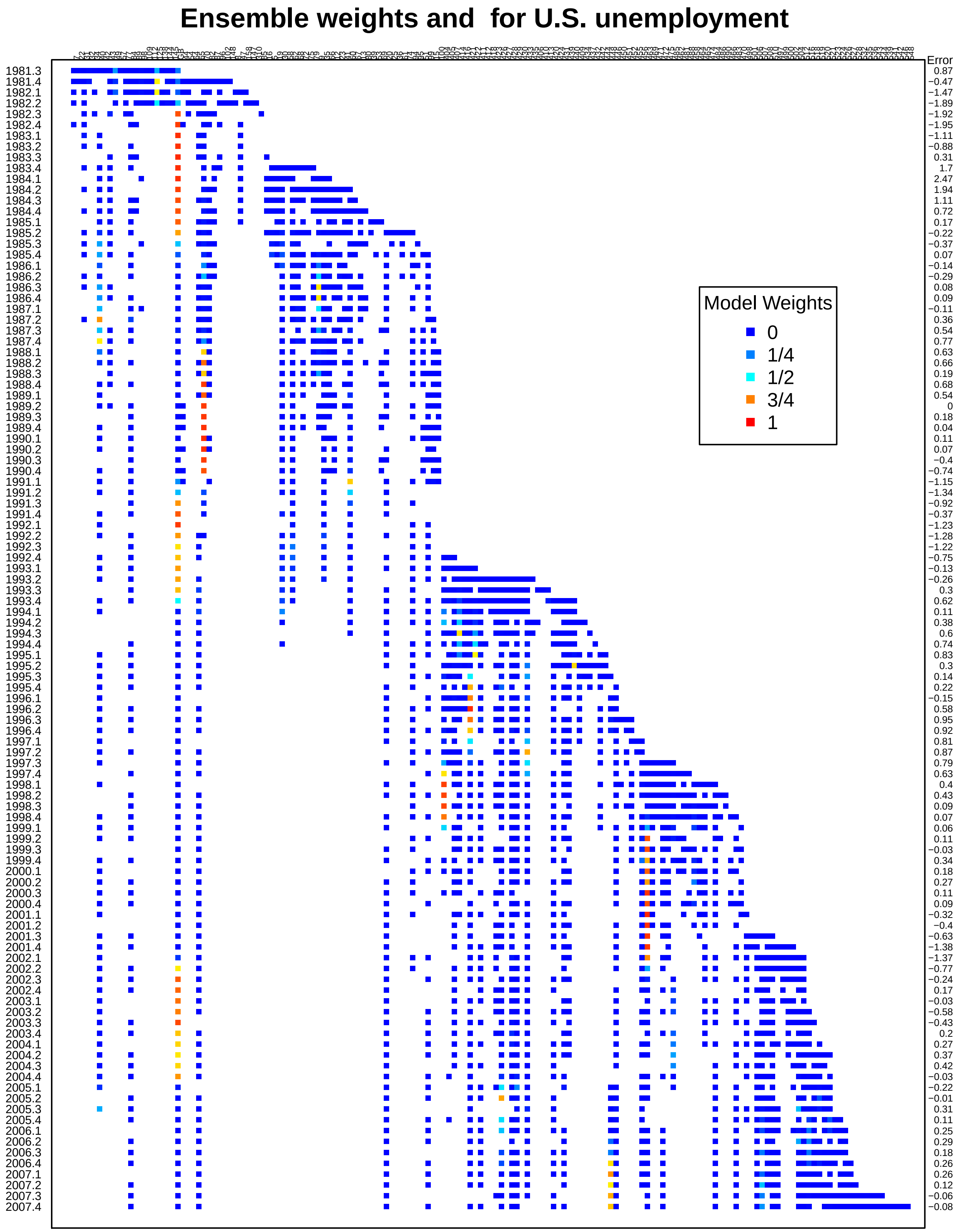
Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui and Michael Polakowski. 2005. "Using Bayesian Model Averaging to Calibrate Forecast Ensembles." *Monthly Weather Review* 133(5):1155–1174.

Vrugt, Jasper A., Cees G.H. Diks and Martyn P. Clark. 2008. "Ensemble Bayesian Model Averaging Using Markov Chain Monte Carlo Sampling." *Environmental Fluid Mechanics* 8(5):579–595.

Wright, Jonathan H. 2008. "Bayesian Model Averaging and Exchange Rate Forecasts." *Journal of Econometrics* 146(2):329–341.

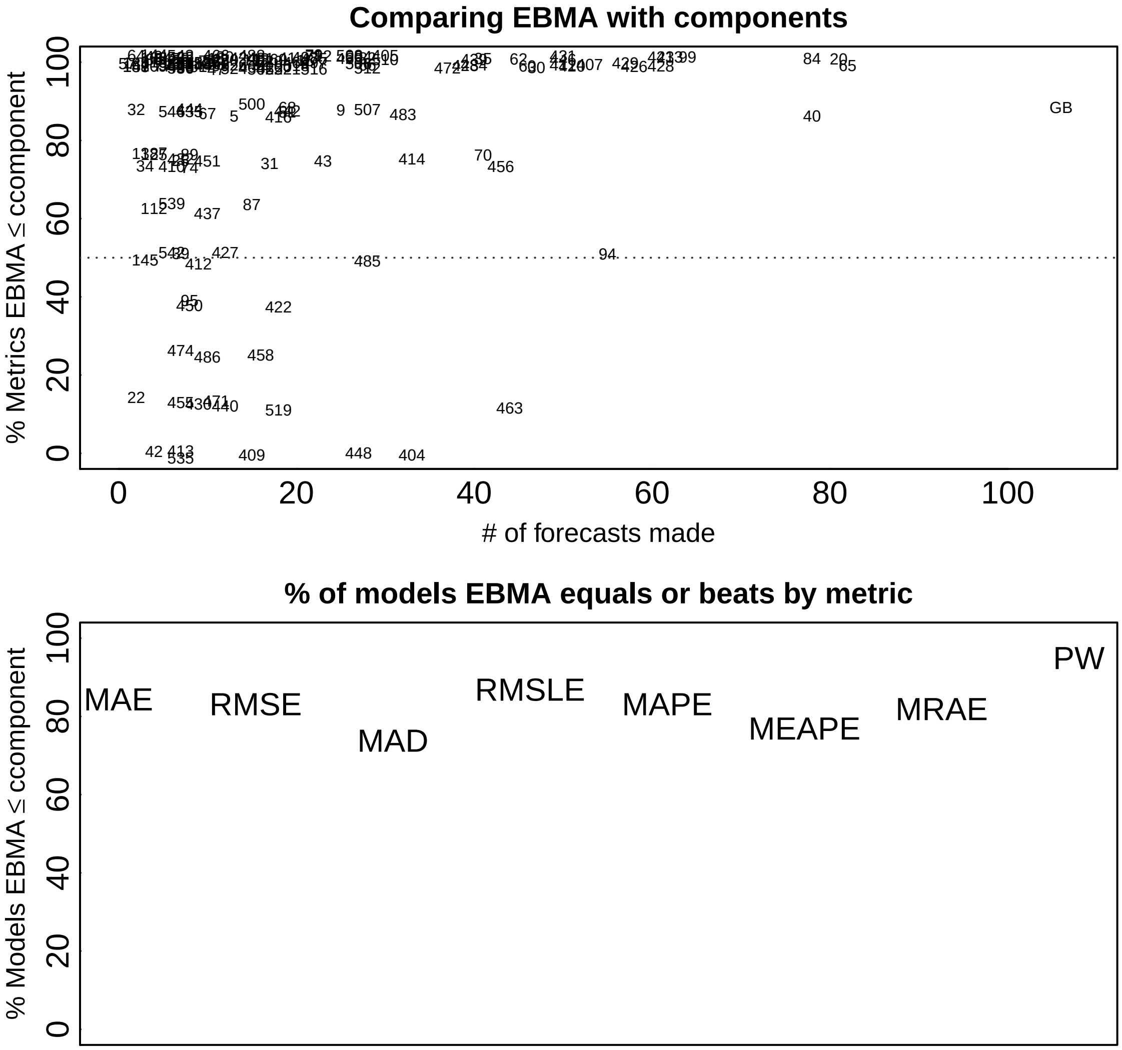
Wright, Jonathan H. 2009. "Forecasting US Inflation by Bayesian Model Averaging." *Journal of Forecasting* 28(2):131–144.

Figure 1: Model weights with rolling EBMA calibration window



This figure shows the component weights for each EBMA model estimated between 1983 and 2007. Ensembles are calibrated on the past ten quarters. The colors going from blue to red indicate increasing weights for components in a given ensemble model. Those components excluded in a given model are left blank.

Figure 2: Performance Comparison



This figure shows the the performance of the EBMA model against its component models on a number of metrics. The top panel shows the percentage of metrics for which EBMA outperforms each of the component models. The bottom panel shows the percent of component models that are beat or tied by the EBMA model for any given metric.

Figure 3: Time-series plot of the Unemployment Forecasts

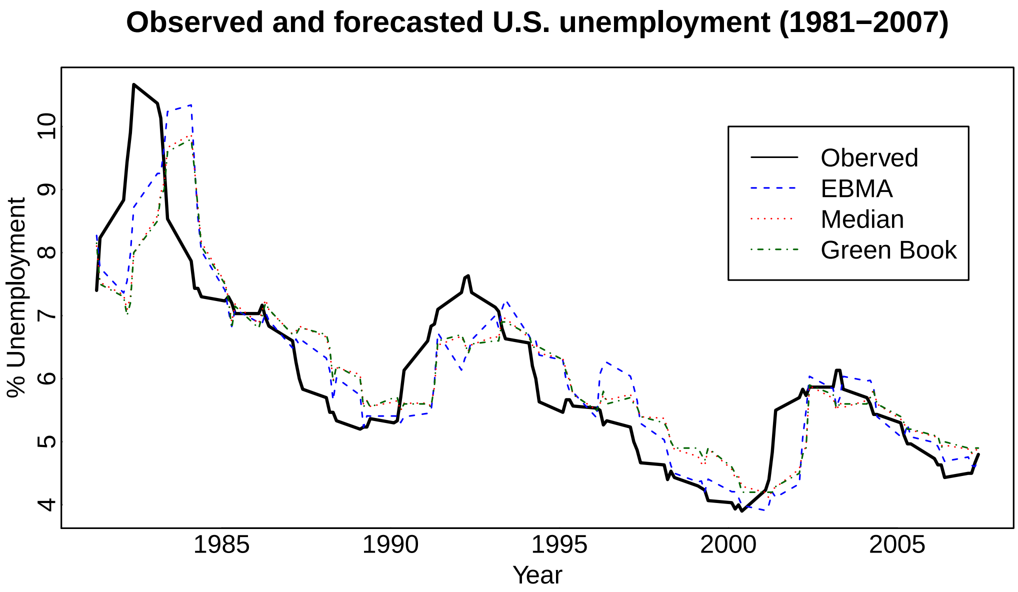
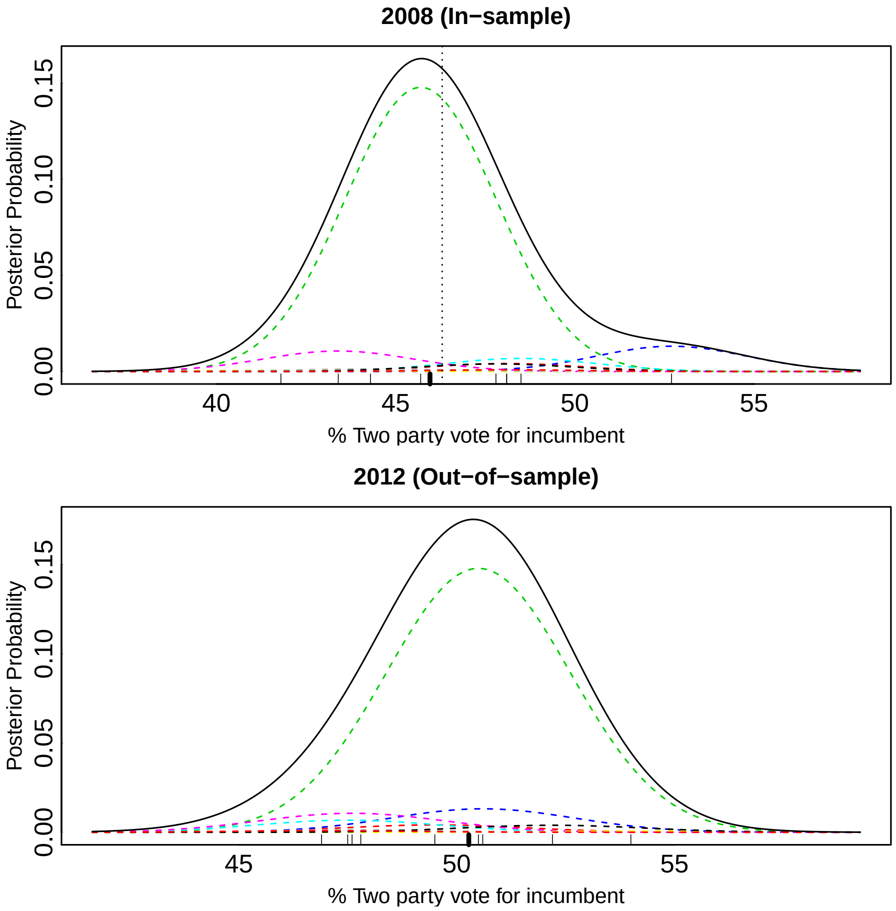


Figure 4: Clever caption here



The figure shows the density functions for each of the component models in different colors and scaled by their respective weight. The point predictions of the individual models are depicted by smalls vertical dashes. The black curve is the density of the EBMA prediction, with the bold dash indicating the EBMA point prediction. For 2008 the vertical dashed line shows the actual result.