Say Yes to the Guess:
Tailoring Elegant 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 14, 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 (EBMA) in the context of small-n prediction tasks with high rates of missing component forecasts. With a large number of observations to calibrate ensembles and low rates of missing values for each component model, the standard approach to calibrating ensembles introduced by Raftery et al. (2005) performs well. However, data in the social sciences generally do not fulfill these requirement. The number of outcomes being predicted tend to be relatively small and missing predictions are neither random nor rare. In these circumstances, EBMA models may overweight components with low rates of missingness and those that that perform well on the limited calibration sample. This can seriously undermine the advantages of the ensemble approach to prediction. We demonstrate this problem and provide a solution that diminishes these undesirable outcomes by introducing a "wisdom of the crowds" parameter to the standard EBMA framework. We show that this solution improves predictive accuracy of EBMA forecasts in both political and economic applications.

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 public who are 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 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, scholars have shown ensemble predictions to be more accurate than any individual component model and 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 Hastie, Tibshirani and Friedman (2009). A wide range of statistical approaches including neural nets, bagging, random forests, additive regression trees, and boosting and more may be properly considered ensemble approaches.

One such method advocated recently for forecasting is ensemble Bayesian model averaging (EBMA). This methods 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 predictive probability density function (PDF). EBMA mixture models seek to collate the good parts of existing forecasting models, while 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 PDF.

However, there are several challenges for creating ensemble predictions for many application 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* provides only 20.

A second and related issue is that in many forecasting exercises there tend to be many 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 forecast for various economic quantities. An 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, there are likely to be many missing forecasts in any given time window. 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 and some forecasting efforts are only active for short time-periods due to poor performance. Moreover, forecasts have tended to accumulate with more potential components 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 the 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 forecasting models. Moreover, several of the individual forecasts are missing for a significant portion of the data. 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 election forecasting, these data issues are endemic

[1] See, for example Fair (2009, 2011); Abramowitz (2008); Campbell (2008); Hibbs (2012); Lockerbie (2008); Erikson and Wlezien (2008); Graefe et al. (2010); Holbrook (2008). 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 use the in-sample fitted values of the election forecasting models 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.

to the social sciences. 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 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 to ensure that no one component is overweighted due to strong performance in the limited calibration period. Below, we first introduce the basic EBMA model in section 2. We 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 the presidential election forecasts shown in Table 1.

2 Notation and basic EBMA model

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 have a prior probability distribution $M_k \sim \pi(M_k)$. 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 $p(M_k|\mathbf{y}^t) = p(\mathbf{y}^t|M_k)\pi(M_k)/\sum_{k=1}^{K} p(\mathbf{y}^t|M_k)\pi(M_k)$, and the and the marginal predictive PDF is $p(\mathbf{y}^{t^*}) = \sum_{k=1}^{K} p(\mathbf{y}^{t^*}|M_k)p(M_k|\mathbf{y}^t)$. This latter 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 period T.

2.1 Dynamic ensemble forecasting

The EBMA procedure assumes K forecasting throughout the training (T') calibration (T) and test (T^*) periods. The component models are calibrated in the training period T'. Optimally, the component model predictions for the calibration period T are then 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 $p(\mathbf{y}|\mathbf{f}_1^{s|t}, \ldots, \mathbf{f}_K^{s|t}) = \sum_{k=1}^{K} w_k g_k(\mathbf{y}|\mathbf{f}_k^{s|t})$, 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 $p(y|f_1^{t^*}, \ldots, f_K^{t^*}) = \sum_{k=1}^{K} w_k g_k(y|f_k^{t^*})$.

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 means are determined by $f_k^{t^*}$ and which are 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 methods. 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 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

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

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 method, however, requires all component models to make predictions for all observations in the calibration period. Thus, if one model's observation is missing at time t, the only solution is listwise deletion. To better accommodate missing values in component models predic-

tion 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 $\tilde{z}_k^{(j+1)t} \equiv \sum_{k \in A^t} \hat{w}_k^{(j)} p^{(j)}(y|f_k^t) / \sum_{k \in A^t} w_k^{(j)}$. 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

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

and

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

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

[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).

4 Small sample adjustment

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 component model predictions.

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 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 $c = 1$, 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 in Montgomery, Hollenbach and Ward (2012) represents a special case of this more general model where $c = 0$.

5 Applications

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 party in U.S. presidential elections.

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

5.1 Quarterly unemployment

Forecasting macroeconomic variables is a quite common exercise in the field of economics and statistics. Calculating accurate forecasts of economic variables is a necessity for policy makers and businesses. These forecasts are created using a wide variety of statistical models.[5] The majority of scholars employ time-series models, with the most commonly applied statistical method being autoregressive integrated moving average (ARIMA) and vector autoregressive (VAR) models. The sophistication and complexity of forecasting models has increased considerably over time. In particular, non-linear dynamic models have gained prominence including threshold autoregressive models, Markov switching autoregressive models and smooth transition autoregression (Elliott and Timmermann 2008; Montgomery et al. 1998). More recently, forecasters have introduced Bayesian VAR models and state-space models to their arsenal (De Goijer and Hyndman 2006; Elliott and Timmermann 2008).

Unsurprisingly, given the large number of ongoing forecasts, scholars have attempted to improve predictive accuracy by combining forecasts Bates and Granger (1969); Palm and Zellner (1992); Elliott and Timmermann (2008). Recently, EBMA and related Bayesian model averaging methods have been successfully employed to create ensemble forecasts of various macroeconomic indicators including inflation (Koop and Korobilis 2009; Wright 2009), GDP (Billio et al. 2010), stock prices (Billio et al. 2011), and exchange rates (Wright 2008).

Policy makers too have come to rely on ensemble forecasts of a sort. The desire to aggregate the collective wisdom of multiple forecasting temas is apparent in the *Survey of Professional Forecasters (SPF)* published by the *Federal Reserve Bank of Philadelphia*. The *SPF* includes forecasts for a large number of macroeconomic variables in the U.S., including the unemployment rate, inflation, and GDP growth.[6] In the first month of every quarter, a survey is sent to selected fore-

[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).

[6]The *SPF* was first administered in 1968 by the American Statistical Association and the National Bureau of Economic Research (NBER). Since 1990, however, it is administered by the Federal Reserve Bank of Philadelphia. `http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/`

casters and is returned by the middle of the second month of the quarter. Forecasts are made for the current quarter as well as several quarters into the future.

This plethora of predictions seems ideal for applying EBMA. Nonetheless, it is plagued by the same issues as discussed for the presidential election forecasts shown in Table 1. Even for quarterly measures, there are relatively few observations, many forecasting teams, and a significant number of missing observations. This setting, therefore, provides an ideal test bed for the adjusted EBMA model discussed above.

Here we focus on forecasting the civilian unemployment rate (UNEMP) as published by the *SPF*. For this application, we selected 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, 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 are missing.[7]

To provide a meaningful benchmark, we also include the "Green Book" 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 (FOMC).[8]

Taking the *SPF* and Green Book unemployment forecasts, we calibrate an ensemble model for each period t, using forecaster performance over the past ten quarters. Only forecasts that had made predictions for five of these quarters were included in the ensemble. 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 Green Book forecasts are sequestered for five years, we generate forecasts beginning in the third quarter of 1983 and running through the fourth quarter of 2007.

[7] On average only 8.4 per cent of all teams make a forecast for any one quarter.

[8] To evaluate forecasts we use the most recent vintage available. All predictions are evaluated using the historical unemployment rate for each quarter. 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 component forecast models. However, in a future version of this paper we will replicate this analysis using real-time data to evaluate the forecasts.

Figure 1 provides a visual representation of EBMA model calibrations throughout this period. In this figure, the wisdom of crowds tuning parameter is 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 blank). Models assigned no weight 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. For any given year, only a subset of forecasting teams offer a prediction. Further, an even smaller subset contains models that both offer a predictions and have made a sufficiently large number of prior forecasts to facilitate 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. Further details for these metrics are shown in the Appendix (Brandt, Freeman and Schrodt 2011).

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 Green Book (GB). It is also worth noting that only 6 out of the total 163 components outperforms 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 Green Book, (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.

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. **?** 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 forecast is closer to the actual rate than the median SPF or the Green Book forecast.

Table 2: Comparing adjusted EBMA models with Green Book, median, and mean forecasts of U.S. Unemployment (1981-2007)

	MAE	RMSE	MAD	RMSLE	MAPE	MEAPE	MRAE	PW
EBMA (c=0)	0.54	0.74	0.37	0.093	8.37	6.49	**0.73**	**27.36**
EBMA (c=0.05)	**0.54**	0.74	**0.37**	**0.093**	**8.33**	**6.30**	0.75	**27.36**
EBMA (c=0.1)	0.54	0.74	0.35	0.093	8.40	6.44	0.76	28.30
EBMA (c=1)	0.61	0.80	0.46	0.102	9.72	8.92	0.95	46.23
Green Book	0.57	**0.73**	0.43	0.093	9.37	8.81	1.00	45.28
Forecast Median	0.62	0.81	0.47	0.103	9.83	8.87	0.98	47.17
Forecast Mean	0.61	0.80	0.46	0.102	9.71	9.06	0.93	46.23

Definitions of model fit statistics are provided in the Appendix. The model with the lowest score for each metric are shown in bold.

[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 $c = 1$.

Table 2 formally 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 Green Book outperforms the ensemble by 0.01 on RMSE), the EBMA model outperforms both the Green Book forecast and the unweighted mean and median forecast on every metric. 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.

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

Table 3: Model weights and in-sample fit statistics for EBMA model of U.S. Presidential Elections (1992-2008)

	EBMA Weight	RMSE	MAE
EBMA		1.92	1.56
Fair	0.02	5.53	4.58
Abramowitz	0.78	2.02	1.72
Campbell	0.07	3.46	2.88
Hibbs	0.04	2.68	2.44
Lewis-Beck, Rice, and Tien	0.06	2.78	2.28
Lockerbie	0.00	7.33	6.97
Holbrook	0.01	5.73	4.77
Erikson and Wlezien	0.02	2.74	2.25
Cuzàn	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, 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 listwise 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 reviewing the standard EBMA framework, 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 minimal 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 adjusted EBMA model in 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 Green Book. 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 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 (the uncertainty estimated around the prediction is quite large).

Finally, we note that in future drafts of this paper, we hope to (1) compare alternative methods of handling missing data (2) discuss how to select window of time for calibration and (3) conduct some simulation studies to explore settings for c parameter and to test the numerical stability of our results.

Predictive Metrics Appendix

Denote the forecast of observation i as f_i and the observed outcome as y_i. We define the *absolute error* as $e_i \equiv |f_i - y_i|$ and the *absolute percentage error* as $a_i \equiv e_i/|y_i|$. Finally, for each observation we have prediction from naive forecast, r_i, that serves as a baseline for comparison. In the example is the main text, this naive model is simply the lagged observation. We can therefore define $b_i \equiv |r_i - y_i|$.[10]

Donating the median of some vector $\mathbf{x}$ as $med(\mathbf{x})$, and the standard indicator function as $I(.)$,

$$MAE = \frac{\sum_1^n e_i}{n},$$

$$RMSE = \sqrt{\frac{\sum_1^n e_i^2}{n}},$$

$$MAD = med(\mathbf{e}),$$

$$RMSLE = \sqrt{\frac{\sum_1^n \left(ln(f_i + 1) - ln(y_i + 1)\right)^2}{n}},$$

$$MAPE = \frac{\sum_1^n a_i}{n},$$

$$MEAPE = med(\mathbf{a}),$$

$$MRAE = med\left(\frac{e_1}{b_1}, \ldots, \frac{e_n}{b_n}\right),$$

and

$$PW = \frac{\sum_1^n I(e_i > b_i)}{n}$$

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*

[10] See Brandt, Freeman and Schrodt (2011) for additional discussion of comparative fit metrics.

Researchers and Practitioners, ed. J. Scott Armstrong. Norwell, MA: Kluwer Academic Publishers.

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).

Brandt, Patrick T., John R. Freeman and Philip A. Schrodt. 2011. "Racing Horses: Constructing and Evaluating Forecasts in Political Science." Paper prepared for the 28th Annual Summer Meeting of the Society for Political Methodology. `http://polmeth.wustl.edu/media/Paper/RHMethods20110721small_1.pdf` (accessed August 20, 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.

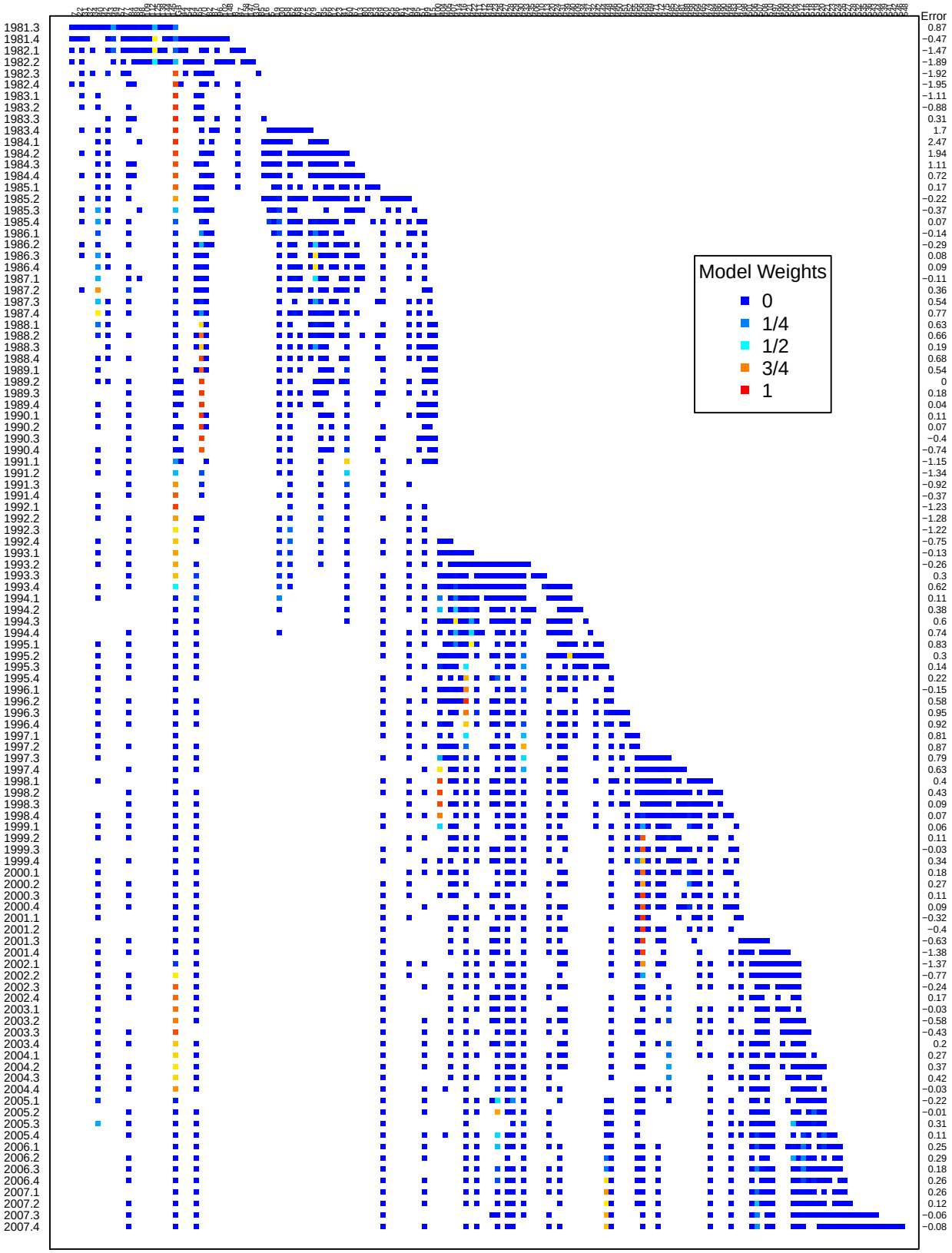
Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui and Michael Polakowski. 2005. "Using Bayesian Model Averaging to Calibrate Forecast Ensembles." *Monthly Weather Review* 133(5):1155–1174.

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.

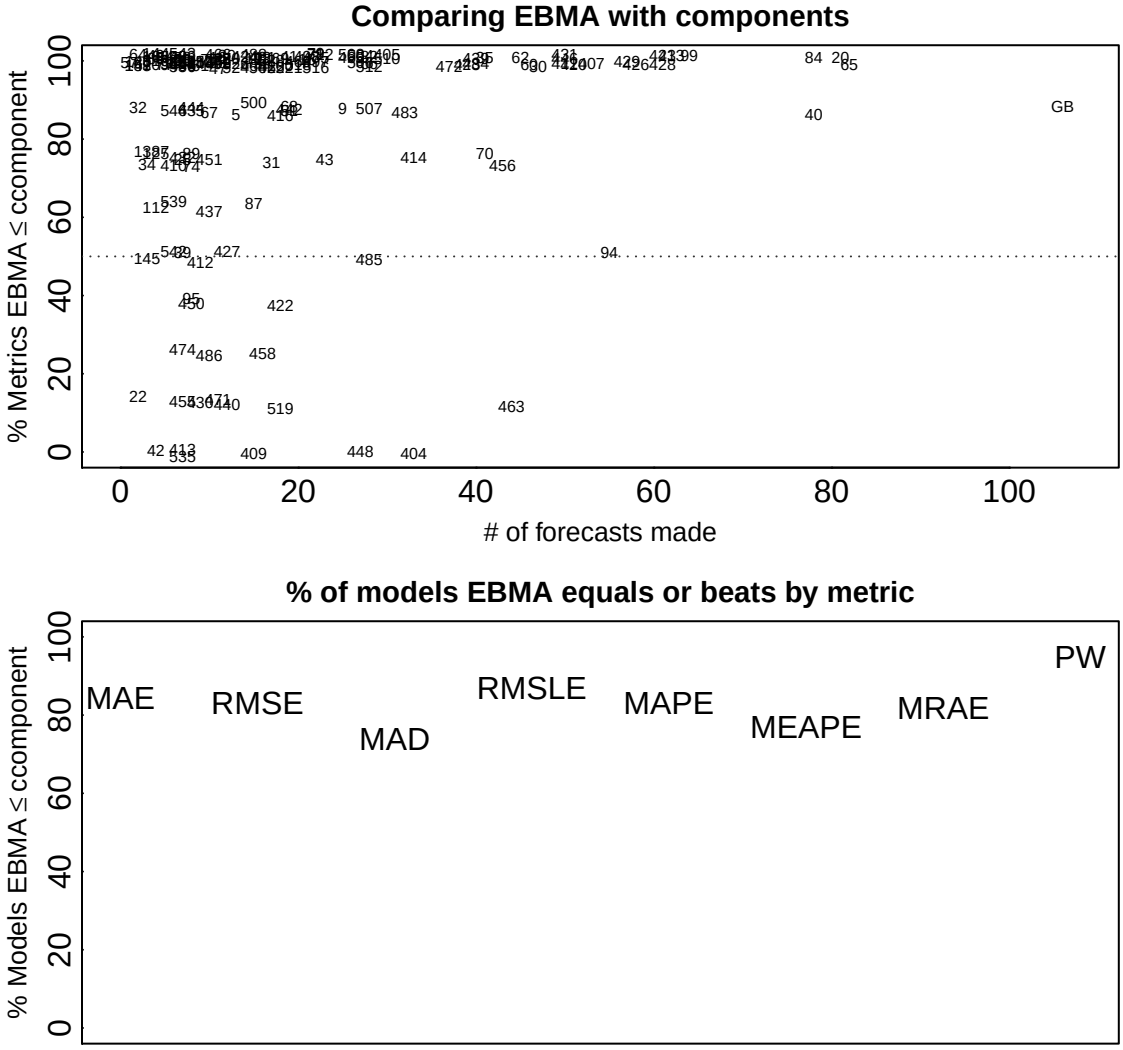
Wright, Jonathan H. 2009. "Forecasting US Inflation by Bayesian Model Averaging." *Journal of Forecasting* 28(2):131–144.

Figure 1: Ensemble weights for SPF forecasts of U.S. unemployment with a rolling 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

Figure 2: Comparing predictive accuracy of EBMA and component models with eight metrics



The top panel plots EBMA's relative performance, as measured by eight forecasting metrics, to each of its components against the number of forecasts generated by the component models. The bottom panel shows the percentage of component models that EBMA matches or outperforms as measured by each metric. Details on the eight forecasting metrics are shown in the Appendix.

Figure 3: Observed and forecasted U.S. unemployment (1981-2007)

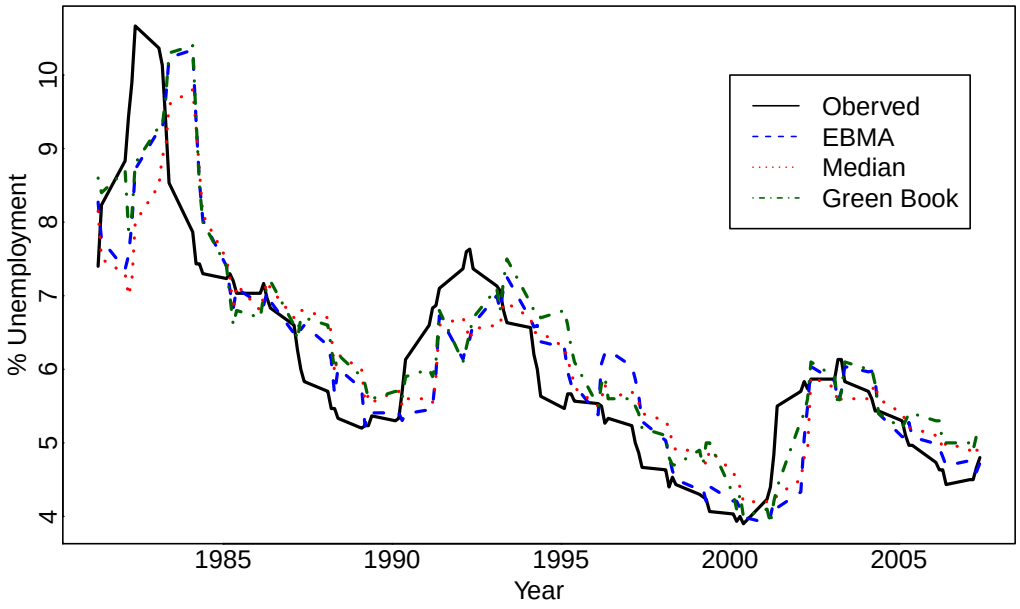
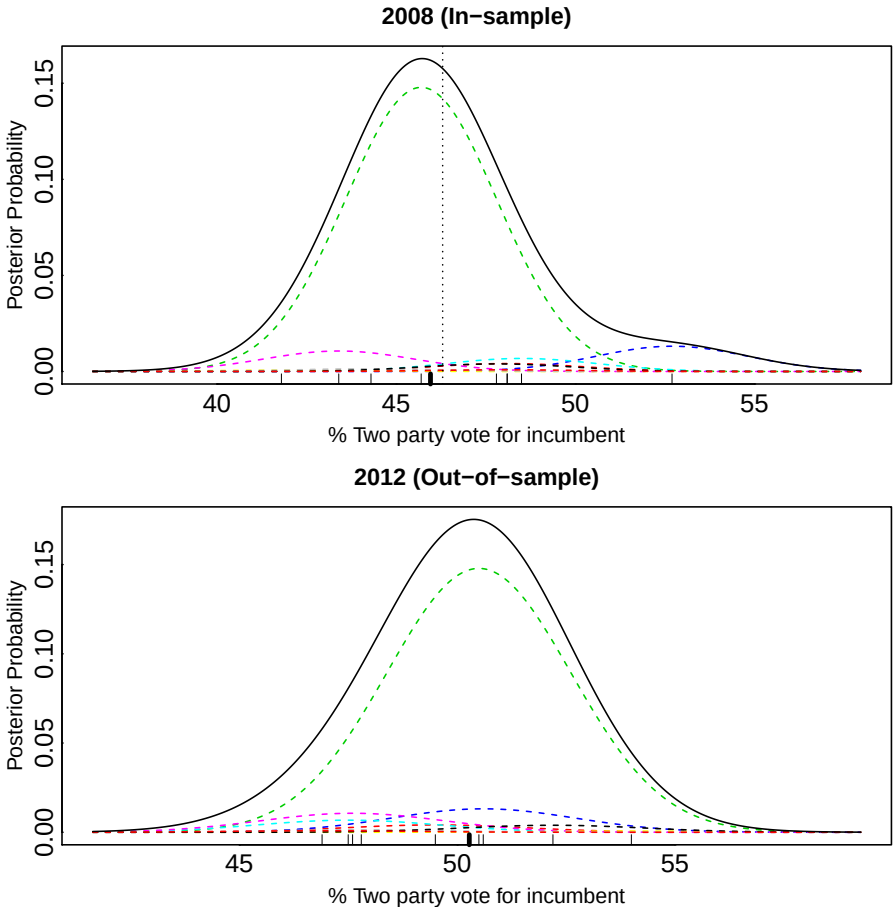


Figure 4: Predictive ensemble PDFs of incumbent-part vote share in U.S. Presidential Elections



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.