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 10, 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 only a few ensembles are missing estimates the standard approaches introduced by Raftery et al. (2005) work fine. If, however, ensembles 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 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, and boosting and more may be properly considered ensemble approaches.

One ensemble method advocated recently for forecasting is ensemble Bayesian model averaging (EBMA). This methods was first proposed by Raftery et al. (2005) and recently forwarded as a useful method for the social sciences by Montgomery, Hollenbach and Ward (2012). In essence, EBMA creates a finite mixture model that creates a kind of weighted average of forecasts. EBMA mixture models seek to collate the good parts of existing forecasting models while avoiding overfitting 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.

However, there are several challenges for creating ensemble predictions for many social science applications. To begin with 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.

A second and related issue is that there tends 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 forecast for various economic quantities. An example are the various forecasts of the FOMC of the U.S. Federal Reserve Board, which are frequently updated, as for example in `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. 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. 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 entirety of scholarly forecasts which produced more than one forecast for elections in the 20th century [1]. In this instance we have only five observations by which to calibrate a model while we have nine forecasts. Moreover, several of the individual forecasts are missing for a significant portion of the data. The forecast of Cuzan, for instance, is missing for 60% of the elections in the dataset.

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	47.7

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**an. Data taken from the collation presented at `http://fivethirtyeight.blogs.nytimes.com/2012/08/07/models-models-everywhere/`.

While particularly egregious for presidential forecasting, these data issues are endemic across 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

[1](Fair 2009, 2011; Abramowitz 2008; Campbell 2008; **?**; Lockerbie 2008; Erikson and Wlezien 2008; Graefe et al. 2010; Holbrook 2008, See, for example). A recent symposium in *PS: Political Science* presents and summarizes recent attempts by a variety of scholars to predict the 2012 U.S. Presidential Election. Also something here about how we previously calibrated based on in-sample, but here we are focused on pure out-of-sample forecasts

even in the presence of these kinds of data-quality issues. Specifically, we show EBMA can be adjusted to easily accommodate missing forecasts. In addition, we propose an alteration to the basic model. Below, we briefly 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 presidential forecasting models shown in Table reftab:one.

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 $p(M_k|\mathbf{y}^t) = p(\mathbf{y}^t|M_k)\pi(M_k)/\sum\limits_{k=1}^{K} p(\mathbf{y}^t|M_k)\pi(M_k)$ 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 period T.

2.1 Dynamic ensemble forecasting

The EBMA procedure assumes K forecasting throughout the training (T') calibration (T) and test (T^*) periods. 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-process the predictions for out-of-sample bias reduction and treat 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, EBMA model is fully specified by estimating model weights, $w_1, \ldots, w_k$ and the common variance parameter σ^2. We estimate these by 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

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

This 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 validation 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

To accommodate missing ensemble values, (Fraley, Raftery and Gneiting 2010) modify the EM algorithm as follows. [2] 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 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

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

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. Thus, 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 $\sum\limits_{k=1}^{K} z_k^t = 1$, this implies that $z_k^t \in [\frac{c}{K}, (1 - c)]$. To achieve this, we replace Equation 4 above with

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

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

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

period t represents a special case of EBMA where $c = 1$.[3] Likewise, the general EBMA discussed in Montgomery, Hollenbach and Ward (2012) represents 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 in U.S. presidential elections.

5.1 Quarterly unemployment

Blah, blah. Short lit review of this topic here. We are looking at predictions of U.S. unemployment four quarters out.

For each period t, we calibrate an ensemble model 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. In addition to forecasts collected by the survey, we include the "Green Book" forecasts produced by the Federal reserve. This model serves both as a component of the ensemble and as a true baseline model with which to compare the EBMA forecasts. 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.

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 (excluded components are simply blank). In this figure models assigned no weight are shown in dark blue while models that are heavily weighted are shown in red.

Figure 1 shows clearly 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 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 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, the each model k offers forecasts for only a subset of cases $n_k \subset n$. To

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

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. 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.[4] The first three of these forecasts and the true level of unemployment are shown in Figure 3.

Table 2: clever caption here

	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
Green Book	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

The model with the lowest score for each metric are shown in bold.

Table 2 compares these baseline models using all eight of the metrics to EBMA moels 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, EBMA model outperforms both the Green Book forecast and unweighted mean and median forecast. Moreover, these results indicate that the c parameter is best set to a smal 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 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

[4]Note that the EBMA model is calculated 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 $c = 1$.

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: The model names need to be fixed. Nice caption here.

	W	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	0.06	2.78	2.28
Lockerbie	0.00	7.33	6.97
Holbrook	0.01	5.73	4.77
Erikson	0.02	2.74	2.25
Cuzan	0.00	0.99	0.75

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, but that our credible intervals are quite wide.

6 Discussion

Missing data is always a conundrum, and a pain. If there is missing data in ensembles even more pernicious effects are possible as the weights assigned by EBMA typically overweight the ensembles with a lot of missing data, and as a result 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.

JACOB, statisticians will want to know why we don't just estimate C.

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 (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.

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

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.

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.

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

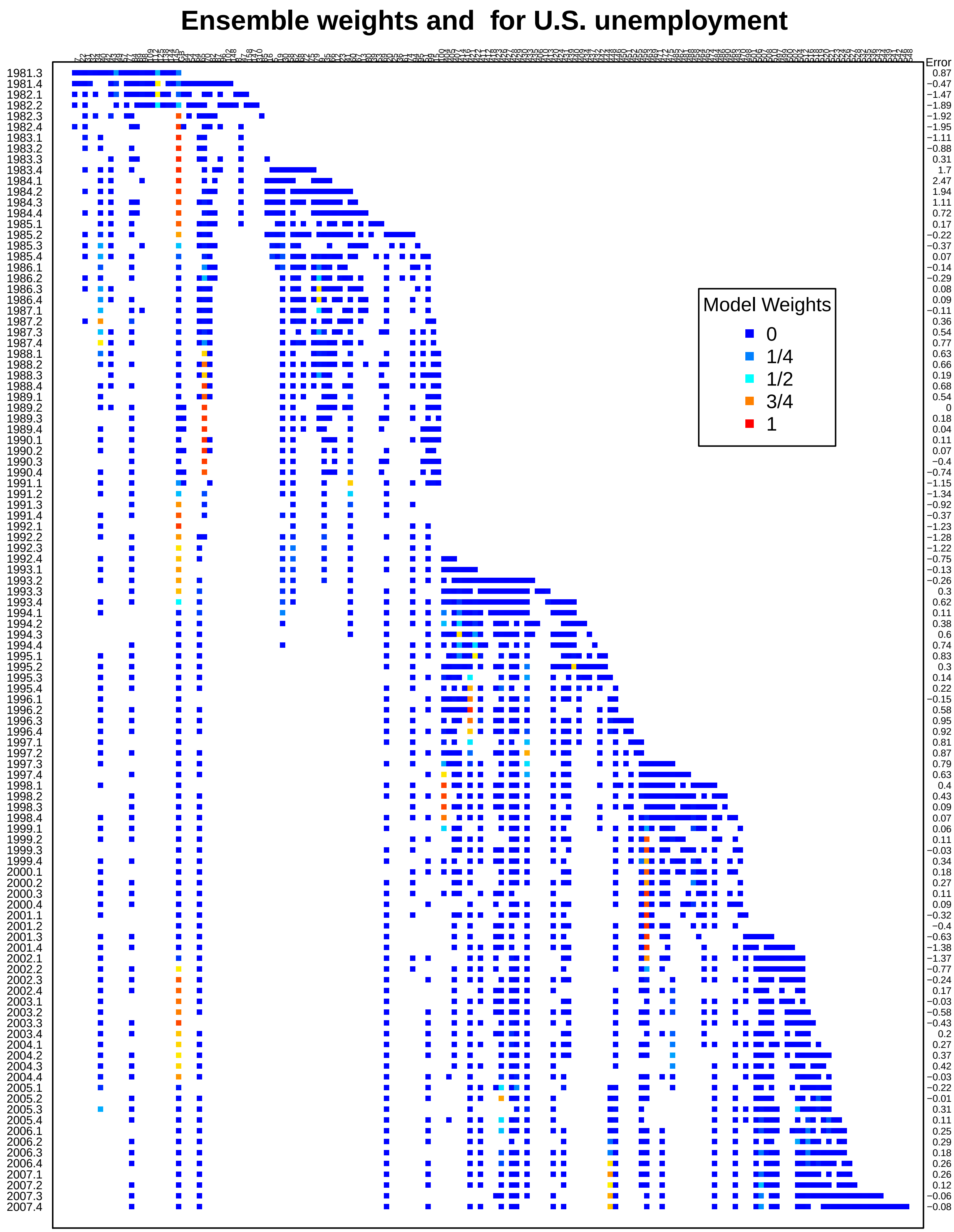
McLachlan, Geoffrey and David Peel. 2000. *Finite Mixture Models*. New York, NY: John Wiley & Sons, Ltd.

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

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.

Figure 1: Clever caption here



Description of figure.

Figure 2: Clever caption here

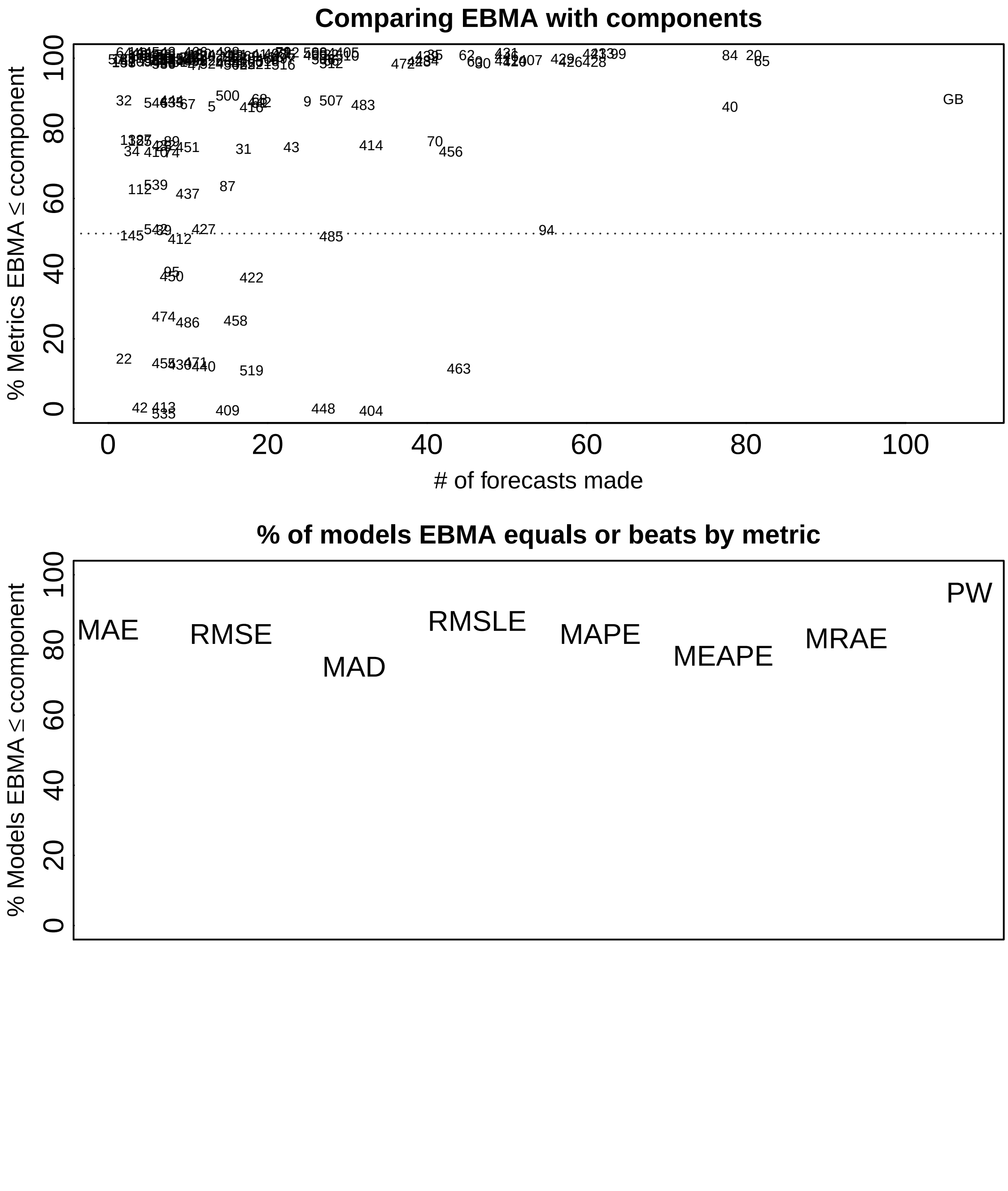


Figure 3: Clever caption here

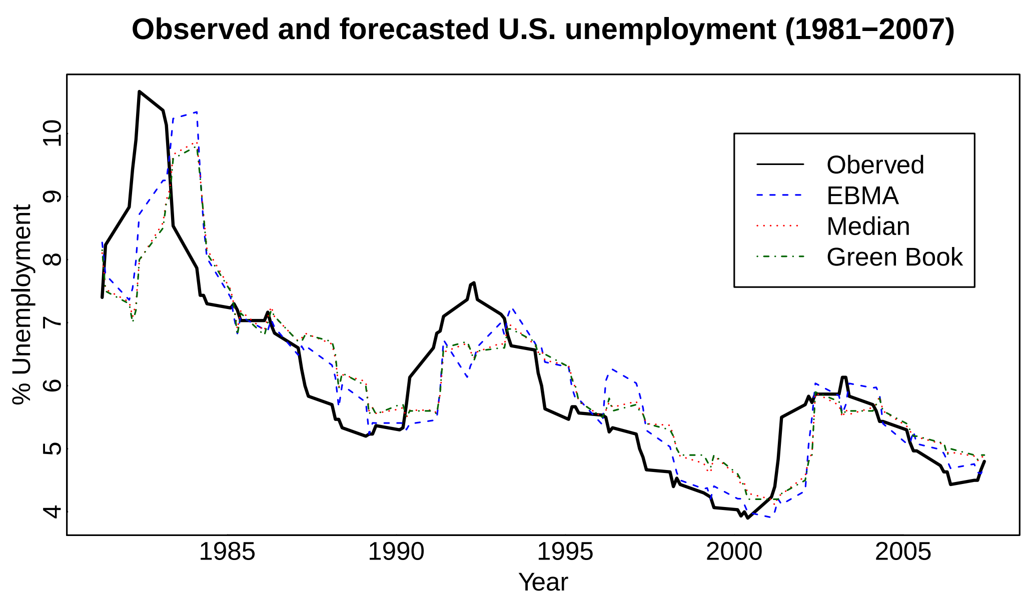
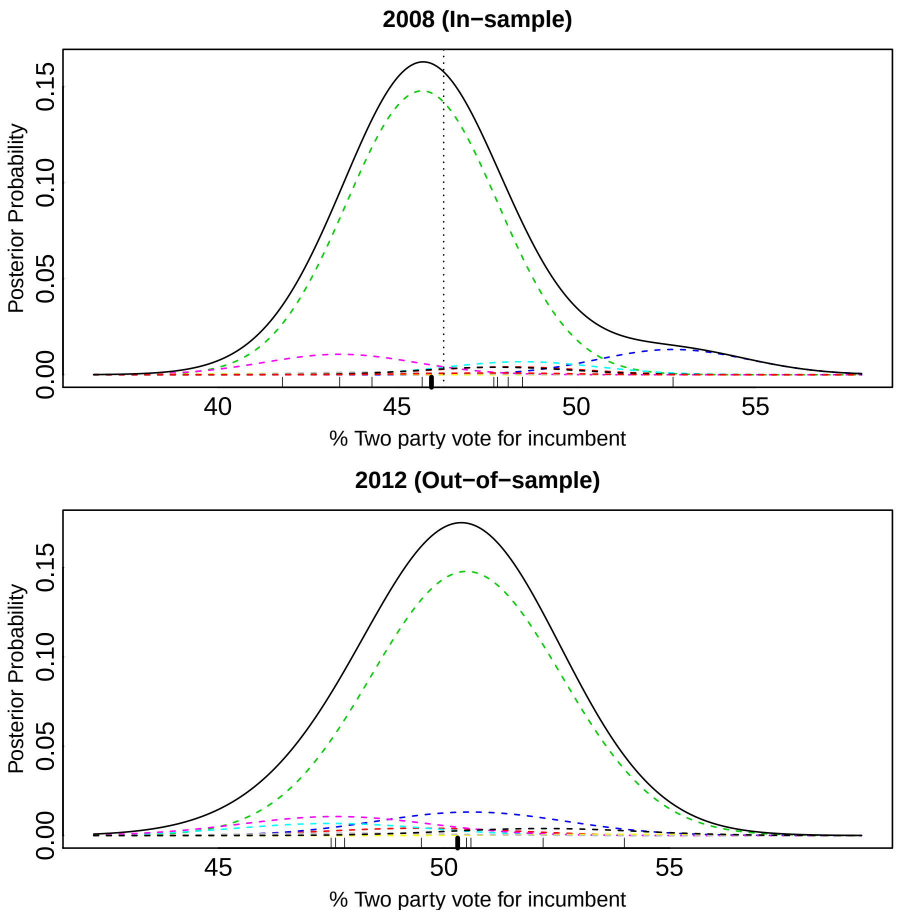


Figure 4: Clever caption here



Explain the figure