Ensemble Predictions of the 2012 US Presidential Elections

Jacob Montgomery, Washington University St Louis   
Florian M. Hollenbach, Duke University   
Michael D. Ward, Duke University

Since at least 1996 political scientists have been generating true out-of-sample predictions of Presidential elections, and since the 2004 Presidential election this journal has presented comparisons of various forecasting models, published prior to the election. The spirit of these symposia is to use the validation provided by correct predictions to claim additional support for specific models. The underlying assertion is that models that are accurate out-of-sample best capture the essential contexts and determinants of elections.

Thus, one aim of this exercise is to develop the “best” model of the underlying data generating process. The main heuristics for comparative evaluation is how well each model does in predicting the electoral results in upcoming election (with some attention given to each model’s inherent plausibility, parsimony, and beauty).

Our approach is entirely different. Rather than search for the best model, theory, or conceptualization of electoral politics, we instead are looking to create the best prediction by combining models. Quite simply we want to make use of the intuition, theories, and concepts implicit in *all* of the forecasting models presented in this symposium to make the most accurate out-of-sample predictions. Without attempting to arbitrate between models and theories, our aim is to aggregate them solely with an eye towards increasing our chances of getting it right.

To do this, without creating a new theory or introducing a new specification, we rely on the extant models. We believe that each of these models captures an important set of insights about US electoral behavior, and each has been rigorously tested not only statistically, but also via a predictive heuristic. Therefore, our approach will attempt to combine the insights of each model into a single predictive ensemble model. It doesn’t matter to our approach if one model “substantively” refutes another. All that matters is that each provides predictions for previous elections that we can use to evaluate their accuracy.

**Ensemble Bayesian Model Averaging**

The approach we use, Ensemble Bayesian Model Averaging (EBMA) was developed recently in the field of weather forecasting as a way of improving predictions by aggregating across models. The core intuition is that there is probably no true “best” model for predicting outcomes of complex systems like the weather. Some weather models might be better at predicting “normal” weather patterns while others are better for rapidly changing conditions. By averaging across multiple prediction models, we can be more accurate, even without having chosen the “best model.”

EBMA uses the predictive performance of the included component models on some *calibration* period to generate a weight for each individual component model. The EBMA prediction is then a kind of weighted average of the predictions made by each of the component models.[[1]](#footnote-1) In particular, the EBMA model will give more weight to models that have been more accurate in the *calibration* period, as well as those models that make more unique predictions.[[2]](#footnote-2) One strong advantage of EBMA is that it only requires each component’s predictions as input, but not the model’s covariates. It is thus possible to use forecasts generated from any kind of process including subject experts, classification trees, or agent based models.

**Mathematical intuition**

More technically, EBMA works in the following way.[[3]](#footnote-3) Assume we have an outcome in the future that is to be predicted and k predictive models (). Each component model comes from a prior distribution and thus one can describe in terms of its probability density function (PDF) conditional on . With the help of simple math and Bayes’ rule, it is then possible to derive the marginal predictive distribution of given the k predictive models as . This PDF can be interpreted as a weighted prediction, where the weights of each model are dependent on the predictive performance in the calibration period prior to t\*.

Applying the EBMA to the example of presidential election forecasting, we use the in-sample predictions of each component model to calibrate the model weights for the EBMA model. Thus we have 12 forecasting models and a training period of 16 presidential elections (from 1948 to 2008). Our test period is the 2012 election.

Each forecasting model is associated with a probability density function (PDF), which is in our case a normal density function centered at the individual forecast . The predictive distribution for observation y2012 (or our forecast for 2012) can be represented as,

where represents the weight associated with each component model. Weights are estimated using maximum likelihood methods.[[4]](#footnote-4) Given the estimated model weights for the EBMA model based on the training period, we can then create an EBMA prediction using the component model predictions.

The basic insight of this approach is that each component model in the ensemble captures some insight that yields predictions that are selectively accurate. Combining them and weighting them by their predictive success creates a sort of meta-model that in principle should be as good (in terms of predictive accuracy) as any individual component model. Across many elections, it is likely that the ensemble will actually dominate each of its members. Indeed, the method has been successfully applied in a wide variety of settings such as inflation (Wright, Forecasting US Inflation by Bayesian Model Averaging, 2009; Tilmann Gneiting, 2010; Gary Koop, 2009), economic growth (Monica Billio, 2010; William A. Brock, 2007), exchange rates (Wright, Bayesian Model Averaging and Exchange Rate Forecasts, 2008), industrial production (Feldkircher, 2010), and weather (Richard M. Chmielecki, 2010; Raftery, 2005; Veronica J. Berrocal, 2010). Its theoretical underpinnings, as well as its success in a variety of contexts, suggest it could be useful in predicting elections as well.

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 1: Ensemble weights and fit statistics for calibration-period performance (1948-2008)* | | | |
|  | Ensemble Weight | RMSE | MAE |
| Ensemble | NA | 1.02 | 0.808 |
| Abramowitz | NA | NA | NA |
| Berry | 0.043 | 1.260 | 1.099 |
| Campbell 1 | 0.284 | 1.610 | 1.252 |
| Campbell 2 | 0.086 | 1.740 | 1.412 |
| Cuzán 1 | 0.035 | 2.300 | 1.751 |
| Cuzán 2 | 0.406 | 1.960 | 1.426 |
| Erikson-Wlezien | 0.026 | 1.780 | 1.568 |
| Hibbs | 0.000 | 2.810 | 2.240 |
| Holbrook | 0.121 | 2.140 | 1.734 |
| Lockerbie | 0.000 | 3.940 | 3.329 |
| *The first column shows the weight assigned each component model in the final ensemble. The other columns show two fit statistics to evaluate the relative performance of each component model and the ensemble across the calibration period. EBMA tends to place higher weight on better performing models, but the relationship is not linear.* | | | |

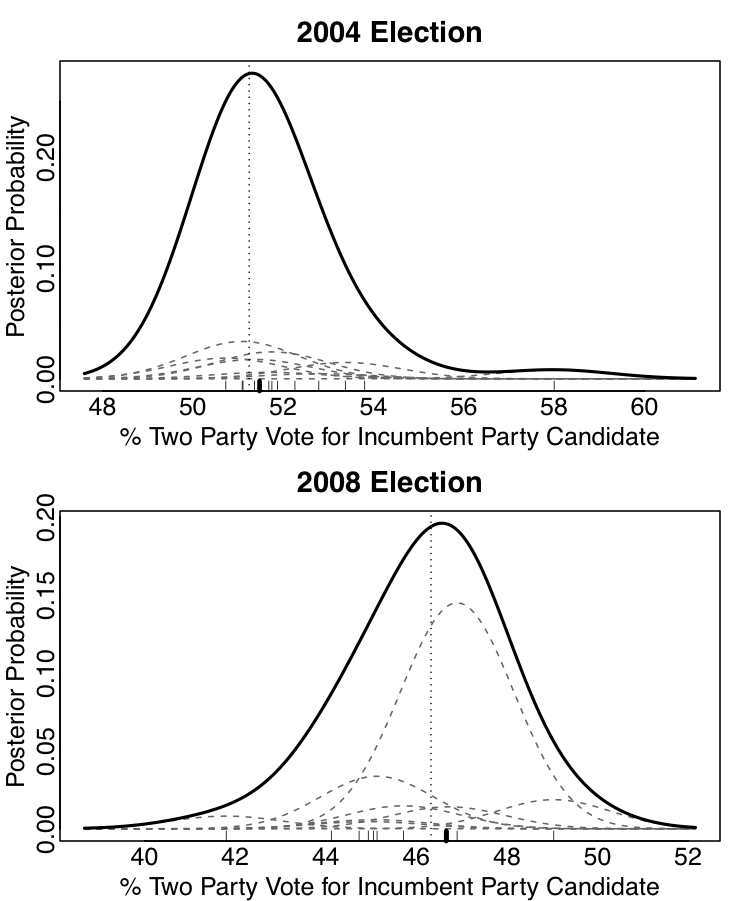
Table 1 shows the EBMA model statistics for the in-sample period. It shows the estimated weights for each individual model, as well as the Root Mean Squared Error (RMSE) and Mean Average Error (MAE) for the in-sample period spanning the post-war era. Almost all of the component models receive some weight in the final ensemble, although the weights are far from uniform. The highest weighted model is Cuzan’s which is based on the idea that a fiscally expansionary policy, even controlling for Fair's variables measuring economic growth, time in office, and party, costs the incumbent party candidate an average of four percent points of the two-party vote. In contrast, EBMA (almost) entirely excludes the Hibbs and Lockerbie models. This should not be interpreted to indicate that some models are “better”, but only that the EBMA procedure found this mix to provide the highest rate of in-sample predictive accuracy.

What can be said about the results is that the EBMA model is more than 20% more accurate than any single model in terms of root mean squared error (RMSE) and more than a third more accurate than any single model in terms of mean average error.

A visual representation of the kinds of predictive PDFs generated by EBMA is provided in Figure 1. The PDF of our EBMA for 2004 and 2008 (in-sample) are illustrated as bold lines, and the predictive densities of the components of the ensemble are shown as dotted lines. The latter have been scaled by the weighting factors that have been developed through the EMBA procedure. These two graphs illustrate that the EBMA can be seen as a covering distribution, and show that for any given in-sample year, the EBMA does not necessarily produce the predictions closest to the actual result (shown as a vertical dashed line), though often it does. In 2008 the posterior PDF contains a small bump to the high end of the incumbent vote (around 49%) that also gets weighted and integrated into the mix, pushing the posterior **median** close to the actual result.

Having developed this approach, we can take the actual predictions of each model in the ensemble to produce an EBMA forecast using the weights we develop, along with **XXXX (JACOB)?** You can find the component predictions in the other articles in this symposium. Using these with the information our approach develops about their individual accuracies, we estimate that the vote for the Democratic Candidate for the 2012 U.S. Presidential Election will be 51.2% with a 95% Credible Interval of

Figure 1: EBMA posterior distributions for the 2004 and 2008 Elections (in-sample).



The dashed curves show the component PDFs and the solid curve shows the final EBMA PDF. The light dashes at the bottom show the point predictions of each component, the dark dash shows the EBMA posterior median, and the vertical dotted line shows the actual election outcome.

# References (this will change as we add refs.

Feldkircher, M. (2010). Forecast Combination and Bayesian Model Averaging: A Prior Sensitivity Analysis. *Journal of Forecasting, {in press}*.

Gary Koop, D. K. (2009). *Forecasting Inflation Using Dynamic Model Averaging.*

Jacob M. Montgomery, F. M. Hollenbach, and Michael D. Ward, (2012, March 22). Improving Predictions Using Ensemble Bayesian Model Averaging. *Political Analysis* *, 20* (tbd), p. tbd.

Monica Billio, R. C. (2010). *Combining Predictive Densities Using {B}ayesian Filtering with Applications to {US} Economics Data.*

Raftery, A. E. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review* *, 133* (5), 1155-1174.

Richard M. Chmielecki, A. E. (2010). Probabilistic Visibility Forecasting Using Bayesian Model Averaging. *Monthly Weather Review* *, 139* (5), 1626-1636.

Stevens, J. (2012, June 24). Political Scientists Are Lousy Forecasters. *New York Times Sunday Review* , p. SR6.

Tilmann Gneiting, T. L. (2010). *Predicting Inflation: Professional Experts Versus No-Change Forecasts.*

Veronica J. Berrocal, A. E. (2010). Probabilistic Weather Forecasting for Winter Road Maintenance. *Journal of the American Statistical Association* *, 105* (490), 522-2537.

William A. Brock, S. N. (2007). Model Uncertainty and Policy Evaluation: Some Theory and Empirics. *Journal of Econometrics* *, 136* (2), 629-664.

Wright, J. H. (2008). Bayesian Model Averaging and Exchange Rate Forecasts. *Journal of Econometrics* *, 146* (2), 329-341.

Wright, J. H. (2009). Forecasting US Inflation by Bayesian Model Averaging. *Journal of Forecasting* *, 28* (2), 131-144.

About the authors

Jacob Montgomery is an Assistant Professor in the Department of Political Science at Washington University in St. Louis and a Fellow at the Center for Political Economy.  His research is in the areas of political methodology and American politics.   He graduated with a B.A. from Wake Forest University with majors in Political Science and Mathematical Economics.  He earned an M.S. in Statistical Science, as well as his Ph.D. (2011) in Political Science from Duke University.

Florian M. Hollenbach is a PhD student with a focus in Political Economy and Methods. His current interests are the political economy of taxation, redistribution as well as state capacity and revenue extraction in developing countries. Additional interests include applied statistics and economic development.

Michael D. Ward is Professor of Political Science at Duke University. His primary interests focus on developing predictive models of political phenomena. His Ph. D. (1977) is from the home of political science predictions, Northwestern University. Much of his current work focuses on predicting the onset and duration of international crises and domestic events such as rebellion, insurgency, and ethnic and religious violence. In addition he works on link prediction in the context of the dynamic evolution of social networks.

1. Ideally, we would calibrate the ensemble model based solely on out-of-sample predictions. This would prevent reliance on models that over-fit the results from prior elections. For practical reasons, however, this is not possible for this forecast. We feel that the models here are sufficiently parsimonious to ameliorate concerns about over-fitting. Nonetheless, we have taken some additional steps (discussed below) to ensure that EBMA does not excessively over-weight any one model. [↑](#footnote-ref-1)
2. This means that component models with highly correlated predictions will be penalized and receive less weight. In addition, EBMA will assign a higher weight for models with fewer missing values in the calibration period. [↑](#footnote-ref-2)
3. We will describe the mathematical detail behind the EBMA model with few details here. For a more detailed description the reader should consult Montgomery et alia, 2012. [↑](#footnote-ref-3)
4. The procedure for calculating model weights for this application builds on the results in Montgomery et alia (2012) in two ways. First, it has been adjusted to handle missingness in forecasts for the calibration period (CITE: Fraley, Raftery, and Gneiting 2010). Second, we have made adjustments to ensure that EBMA does not place excessive weight on single component. This is done to adjust for the fact that the data in the calibration period is not truly out-of-sample. Roughly speaking, the model assumes there is a minimum probability (1/50) that each observation is “best” represented by each of the models. Additional details are provided in CITE APSA PAPER. [↑](#footnote-ref-4)