Ensemble Predictions of the 2012 US Presidential Elections

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

For over two decades now, political scientists have been using statistical models to generate true out-of-sample predictions of presidential elections. Since the 2004 Presidential election, this journal has presented symposia of the various forecasting models published prior to Election Day. The spirit of this exercise 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 does in predicting the electoral results in upcoming election (with some attention given to the models’ 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 models presented elsewhere in this volume. We believe that each captures an important set of insights about U.S. elections, 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 ensemble prediction. 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) derives the field of weather forecasting as a way of improving predictions by aggregating across models (Raftery 2005). The core intuition is that there is probably no true “best” model for predicting outcomes of complex systems like the weather (or presidential elections). Some weather models might be better at predicting “normal” weather patterns while others are better for rapidly changing conditions. However, by averaging across multiple prediction models, we can be more accurate, even without having chosen the “best model.”

EBMA uses the predictive performance of its component models in some *calibration* period to generate a weight for each model. The EBMA prediction is then a kind of weighted average of the predictions made by each of the components.[[1]](#footnote-1) In particular, the EBMA model will give more weight to components that have been more accurate in the calibrationperiod, 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 (). The probability of each predictive model capturing the true data generating process comes from a prior distribution and 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\*.

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)

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 2009; Gneiting & Thorarinsdottir 2010; Koop & Korobilis 2009), economic growth (Billio et al. 2010; Borck, Brock & West 2007), exchange rates (Wright 2008), industrial production (Feldkircher 2010), and weather (Chmielecki & Raftery 2010; Raftery et al. 2005; Berrocal et al. 2010). Its theoretical underpinnings, as well as its success in a variety of contexts, suggest it could be useful in predicting elections as well.

**The EBMA Forecast for 2012**

Applying the EBMA to the example of presidential election forecasting, we use the calibration-period predictions of each component model to estimate the model weights. In this case, we use the model predictions generously provided by Abramowitz, Berry, Campbell (Trial Heat), Cuzan (name?), Erikson-Wlezien, Hibbs, Holbrook, Lewis-Beck/Tien (Jobs), and Lockerbie, which are all described elsewhere in this symposium. Where a single team offered more than one prediction model, we chose the one with the highest calibration-period performance. Thus, we have 9 forecasting models and a training period of 16 presidential elections (from 1948 to 2008). Our test period is, of course, the 2012 election.

[Table 1 about here]

Table 1 shows the EBMA model statistics for the calibration 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 calibration 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. Alan Abramowitz’s model, which is based on June polling data, 2nd quarter GDP growth, and the presence of a first-term incumbent (adjusting for polarization), receives the lion’s share of the predictive weight. In contrast, EBMA (almost) entirely excludes the Hibbs and Lockerbie models.

These weights 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 calibration-sample predictive accuracy. Indeed, it is notable that the model generally places the greatest weight on models that make use of polling data (e.g., Abramowitz), while it gives much less weight to models that offer no predictions for much of the calibration period (e.g., Berry) or those based on data with an older vintage (e.g., Hibbs).

[Figure 1 about here]

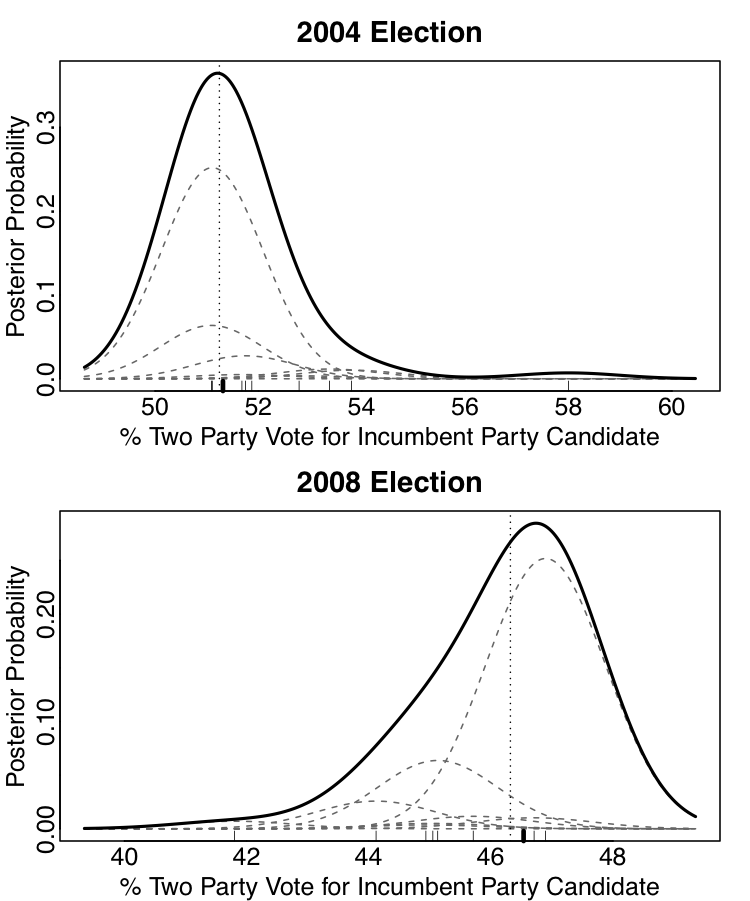
A visual representation of the kinds of predictive PDFs generated by EBMA is provided in Figure 1. The PDF of our EBMA for 2004 and 2008 (in-sample) are illustrated as bold lines, and the predictive densities of the components of the ensemble are shown as dashed lines. The latter have been scaled by the model weights weighting. The point predictions of each component (light dashes) and the ensemble model (bold dash) are shown at the bottom of each plot.

These two graphs show that for any given year, EBMA does not necessarily produce the predictions closest to the actual result (shown as a vertical dotted line), though often it comes very close. However, across many elections EBMA will tend to outperform its component models in terms of accuracy and precision while also accurately reflect the inherent uncertainty implied by the differing predictions coming from the various component models.

With all of this in hand, we finally turn to creating an ensemble forecast for 2012. Using the weights reported in Table 1 and the forecasts provided to us by the respective authors, we estimate that the vote for the Democratic candidate for the 2012 U.S. Presidential Election will be 51% with a 95% credible interval ranging from 48.7% to 53.15%. According to the EBMA posterior, the probability of President Obama winning re-election is 70.37%. Thus, the collective wisdom of this crowd of models -- or at least their wisdom as we have collected them -- is that 2012 will be a close election but that President Obama has a slight edge.

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 1: Ensemble weights and fit statistics for calibration-period performance (1948-2008)* | | | |
|  | Ensemble Weight | RMSE | MAE |
| Ensemble |  | 0.849 | 0.702 |
| Abramowitz | 0.620 | 0.981 | 0.769 |
| Berry | 0.012 | 0.808 | 0.750 |
| Campbell (Trial Heat) | 0.068 | 1.610 | 1.252 |
| Cuzan (name?) | 0.157 | 1.963 | 1.426 |
| Erikson-Wlezien | 0.025 | 1.775 | 1.549 |
| Hibbs | 0.009 | 2.806 | 2.240 |
| Holbrook | 0.029 | 2.144 | 1.734 |
| Lewis-Beck/Tien (Jobs) | 0.063 | 1.264 | 1.050 |
| Lockerbie | 0.017 | 3.943 | 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.* | | | |

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 bolded dash shows the EBMA posterior median, and the vertical dotted line shows the actual election outcome.

# References (this will change as we add refs.

Berrocal, VJ, Raftery, AE, Gneiting, T & Steed, RC 2010, 'Probabilistic Weather Forecasting for Winter Road Maintenance', *Journal of the American Statistical Association*, vol 105, no. 490, pp. 522-2537.

Billio, M, Casarin, R, Van Dijk, HK & Ravazzolo, F 2010, 'Combining Predictive Densities Using

Borck, WA, Brock, N & West, KD 2007, 'Model Uncertainty and Policy Evaluation: Some Theory and Empirics', *Journal of Econometrics*, vol 136, no. 2, pp. 629-664.

Chmielecki, RM & Raftery, AE 2010, 'Probabilistic Visibility Forecasting Using Bayesian Model Averaging', *Monthly Weather Review*, vol 139, no. 5, pp. 1626-1636.

Feldkircher, M 2010, 'Forecast Combination and Bayesian Model Averaging: A Prior Sensitivity Analysis', *Journal of Forecasting*, vol

Fraley, C, Raftery, AE & Gneiting, T 2010, 'Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using

Gneiting, T & Thorarinsdottir, TL 2010, 'Predicting Inflation: Professional Experts Versus No-Change Forecasts'.

Koop, G & Korobilis, D 2009, 'Forecasting Inflation Using Dynamic Model Averaging'.

Montgomery, JM, Hollenbach, FM & Ward, MD 2012, 'Improving Predictions Using Ensemble Bayesian Model Averaging', *Political Analysis*, 22 March 2012, p. tbd.

Raftery, AE, Gneiting, T, Balabdaoui, F & Polakowski, M 2005, 'Using Bayesian Model Averaging to Calibrate Forecast Ensembles', *Monthly Weather Review*, vol 133, no. 5, pp. 1155-1174.

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

Wright, JH 2008, 'Bayesian Model Averaging and Exchange Rate Forecasts', *Journal of Econometrics*, vol 146, no. 2, pp. 329-341.

Wright, JH 2009, 'Forecasting US Inflation by Bayesian Model Averaging', *Journal of Forecasting*, vol 28, no. 2, pp. 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 made in advance of elections. 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. For an introduction to Bayesian model averaging more generally, see Bartels 1997 and Montgomery and Nyhan 2011. [↑](#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 (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 (1/90) that each observation is “best” represented by each of the models. This increases the weight placed on low-probability models and also increases the implied level of uncertainty in the ensemble forecast. Additional details are provided in CITE APSA PAPER. [↑](#footnote-ref-4)