Ensemble Predictions of the 2012 U.S. Presidential Election[1]

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

For over two decades, political scientists have been using statistical models aimed, in part, at making 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 heuristic for comparative evaluation is how well each does in predicting the electoral results in the upcoming election with some attention also given to the models' inherent plausibility, parsimony, and beauty.

[1] We thank Alan Abramowitz, Michael Berry, James Campbell, Alfred Cuzán, Robert Erikson and Christopher Wlezien, Douglas Hibbs, Thomas Holbrook, Michael Lewis-Beck and Charles Tien, as well as Brad Lockerbie for generously sharing their data with us for this enterprise. All their contributions are described in detail elsewhere in this symposium, but are not indexed separately within this manuscript.

Our approach is different. Rather than attempting to create the best model or theory, we instead seek to create an ensemble prediction of the upcoming election. 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 prediction. 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, we rely on the models presented elsewhere in this volume. We believe that each captures an important set of insights about U.S. elections. Our approach combines those insights into a single ensemble prediction. For our purposes, the theoretical differences between the models are irrelevant. All that matters is that each provides predictions for previous elections that we can use to evaluate their accuracy. We then weight each forecast by its previous performance and combine them in an effort to create the most accurate out-of-sample forecast possible that also adequately captures the uncertainty and diversity inherent in these models.

Ensemble Bayesian Model Averaging

The approach we use, ensemble Bayesian model averaging (EBMA), seeks to improve out-of-sample forecasting by aggregating across models (Raftery et al. 2005). The core intuition of EBMA, which originates in the field of weather forecasting, is that there is probably no one "best" model for predicting outcomes of complex systems like the weather (or presidential elections). For instance, 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, forecast accuracy can be improved without having to try to select the "best" model.

EBMA uses the predictive performance of its component models in a *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.[2] In particular, the EBMA model will give more weight to component models that have been more accurate in the calibration period, as well as those models that make more unique predictions.[3] One advantage of EBMA is that it only uses predictions from each component model rather than the full set of covariates and estimated coefficients. It is thus possible to use forecasts generated from any kind of process including subject experts, classification trees, or agent-based models.

Mathematical intuition

[2] 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 because true out-of-sample predictions from the models in this symposium are only available in a relatively small number of cases. For the purposes of this symposium, therefore, we accept these models at face value. While we have taken some additional steps (discussed below) to ensure that EBMA does not excessively over-weight any one model, readers who believe that the component models are over-fit and miss-specified may wish to instead calibrate weights based on true out-of-sample predictions published by authors before each election.

[3] Component models with highly correlated predictions will are penalized and receive less weight. In addition, our EBMA model assigns a higher weight for models with fewer missing values in the calibration period.

More technically, EBMA works in the following way.[4] Assume we have an outcome y^{t^*} in the future that is to be predicted and k predictive models $(M_1, M_2, \dots, M_k)$. The probability of each predictive model capturing the true data-generating process comes from a prior distribution and one can describe y^t in terms of its probability density function (PDF) conditional on M_k. With the help of simple math and Bayes' rule, it is then possible to derive the marginal predictive distribution of y^{t^*} given the k predictive models as $p(y^{t^*}) = \sum_{k=1}^{K} p(y^{t^*} | M_k) p(M_k | y^t)$. This PDF can be interpreted as a weighted prediction, where the weights of each model M_k 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 $N(f_k^{t^*}, \sigma^2)$. The predictive distribution for observation y^{2012} (or our forecast for 2012) can be represented as, $p(y | f_1^{2012}, f_2^{2012}, \dots, f_K^{2012}) = \sum_{k=1}^{K} w_k N(f_k^{2012}, \sigma^2)$,
where w_k represents the weight associated with each component model. Weights are estimated using maximum likelihood methods.[5]

---

[4] We only briefly introduce the mathematical framework for the EBMA model here. For a more detailed description, the reader should consult Montgomery et alia (2012a). For an introductions to the use of Bayesian model averaging in political science, see Bartels (1997), Bartels and Zaller (2001), as well as Montgomery and Nyhan (2010).

[5] The procedure for calculating model weights for this application builds on our earlier results (Montgomery, Hollenbach & Ward 2012a) in two ways. First, it has been adjusted to handle missing-ness in forecasts for the calibration period (Fraley, Raftery & Gneiting 2010). Second, we have made adjustments to ensure that EBMA does not place excessive weight on a single component. This is done because the predictions in the calibration period are not truly out-of-sample. Roughly speaking, the model assumes there is a minimum

The basic insight of the EBMA 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 past predictions creates a sort of meta-model that in principle should yield out-of-sample forecasts that are as accurate as any individual component model in terms of predictive accuracy and precision. 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 2012), 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 empirical contexts, suggest it could be useful in predicting presidential elections as well.

The EBMA forecast for 2012

To apply EBMA to presidential election forecasting, we first use the calibration-period predictions of each component model to estimate the model weights. In this case, we use predictions generously provided by Abramowitz, Berry, Campbell (Trial Heat Model), Cuzán (FPRIME - long), Erikson/Wlezien, Hibbs, Holbrook, Lewis-Beck/Tien (Jobs Model), and Lockerbie from the models described elsewhere in this symposium. The result is an ensemble of nine forecasting models and a training period of 16 presidential elections from 1948 to 2008. Our test period is, of course, the 2012 election.

---

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 Montgomery et al., (2012b) .

[Table 1 about here]

Table 1 shows the EBMA model statistics for the calibration period -- the estimated weights for each individual model, the root mean squared error (RMSE), and the mean average error (MAE) for the calibration period spanning the post-war era. 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 in the election year, 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 while still reflecting a realistic level of predictive uncertainty. 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 measured far in advance of the election (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 PDFs of our EBMA model for 2004 and 2008 (in-sample) are illustrated as bold lines, while the predictive densities of the component models in the ensemble are represented as dashed lines. (The latter have been scaled by the model

weights.) The point predictions of each component (light dashes) and the ensemble model (bold dash) are shown at the bottom of each plot.

These two plots 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 it often comes very close. However, across many elections EBMA will tend to outperform its component models in terms of accuracy while also reflecting the uncertainty implied by the different predictions of the component models.

With our EBMA model in hand, we finally turn to creating our 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 51.0% with a 95% credible interval ranging from 48.7% to 53.2%. According to the EBMA posterior, the probability of President Obama winning re-election is 70.4%. Thus, the collective wisdom of this crowd of models -- or at least their wisdom as we have combined them -- is that 2012 will be a close election but that President Obama has a slight, but non-trivial, edge in terms of the popular vote.

Conclusion

Combining different sets of information is a time-honored tradition of many who forecast important elections. It is now a regular exercise for pundits and scholars to aggregate polls or expert opinions when trying to gain a glimpse into the future. Ensemble methods, as briefly presented and applied here, provide a principled way to weight each component of such aggregations based on accuracy. This approach seeks to collate the good parts of existing models while avoiding over-fitting. The hope is for greater accuracy as both the knowledge and implied uncertainty of a variety of approaches are integrated into

a combined probability distribution. One aspect of our approach that is relatively unique in the extant literatures on presidential vote forecasting (save Bartels 1997 and Bartels and Zaller 2001) is that it accentuates not only our predictions of election outcomes but also our estimate of the uncertainty around those predictions. Neither the literature nor the popular press emphasizes this and too often it is not even reported. Our uncertainty is part of our knowledge, and merits a full reporting and evaluation.

The summer of 2012 has seen an increased level of attention to the ability of political scientists to do "non-lousy" forecasting (Stevens 2012). The irony is that forecasting *per se* has never been at the heart of social science, despite claims to the contrary. It is therefore worth noting that our ensemble depends on the insight contained in all of the individual components, even if their weights may be small. In the end, mere predictive accuracy is not a substitute for substantively oriented research. Our belief is that the act of forecasting serves as an additional heuristic, one normally demanded of scientific endeavors wherein a replication of results in a new situation is expected to produce the same findings. This predictive heuristic enables us to improve our models by showing us where they break down, as well as where they are upheld.

Thus, the predictive enterprise works collectively and individually to improve understanding of the political world. It is important to keep score for both the individual models and the forecasting literature as a whole so we actually know whether or not we are getting better. It is noteworthy that in the one area in which there is some track record of political science forecasting, namely the prediction of the U.S. presidential elections, the accuracy of the well-known models are fairly good by some measures. For example, the average absolute error in-sample is about 1.5%. This is not perfect by any stretch of the

imagination, but hardly lousy. Indeed, it seems clear that not only the diversity of models but also their accuracy has improved over time.

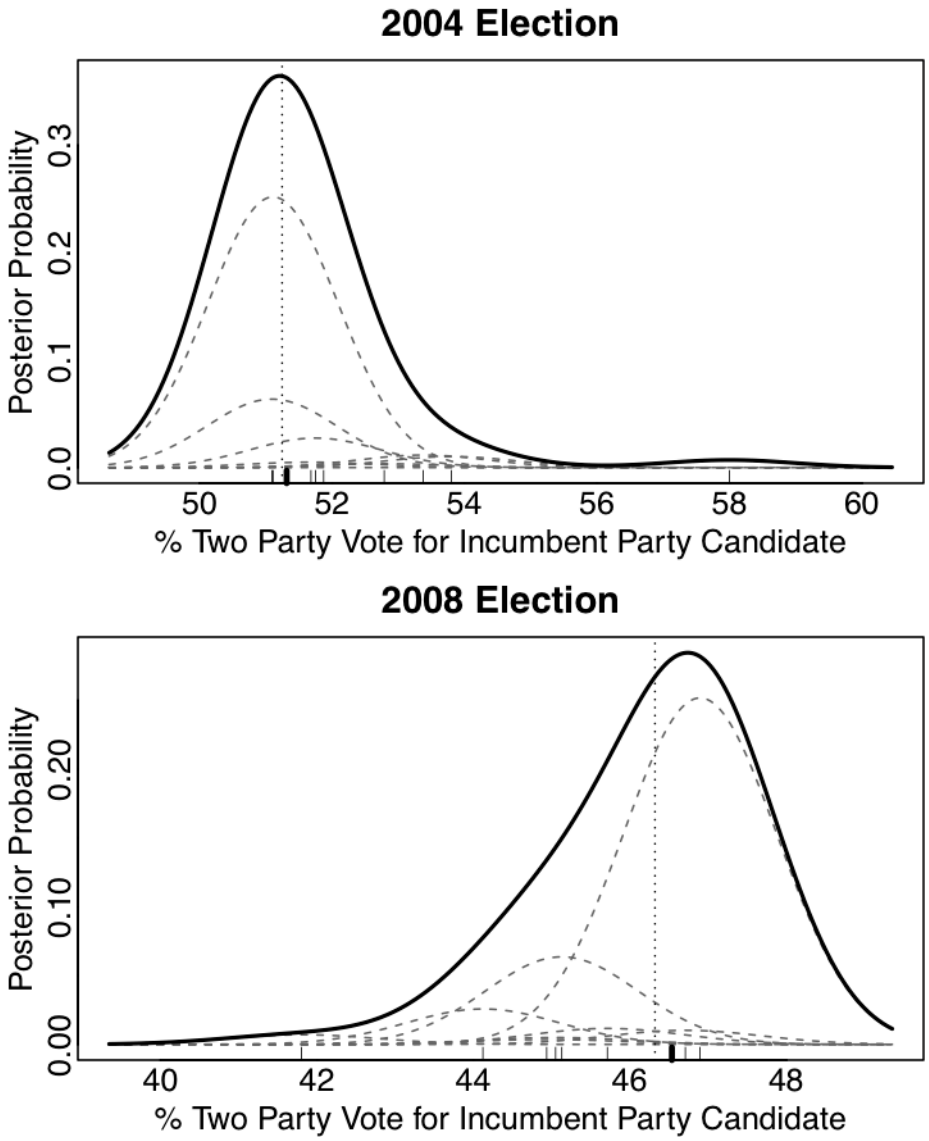
There is no guarantee that our ensemble estimate will be the most accurate in the 2012 election. The future is full of surprises, and we only have a small number of relevant elections on which to construct our ensemble model. However, given what we do know about the performance of these nine models of presidential voting, we can derive an estimate in which we place a high degree of confidence: between 48% and 53% of the U.S. voters will support the incumbent in 2012 and there is a 70% probability that the vote for Obama will be greater than 50%.

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
Cuzán (FPRIME long)	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 second column contains 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

Bartels, L.M., 1997. Specification Uncertainty and Model Averaging. *American Journal of Political Science*, 41(2), pp.641-74.

Bartels, L.M., Zaller, J., 2001. Presidential Vote Models: A Recount. *PS: Political Science and Politics*. 34(1), pp. 9-20.

Berrocal, V.J., Raftery, A.E., Gneiting, T. & Steed, R.C., 2010. Probabilistic Weather Forecasting for Winter Road Maintenance. *Journal of the American Statistical Association*, 105(490), pp.522-2537.

Billio, M., Casarin, R., Van Dijk, H.K. & Ravazzolo, F., 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).

Borck, W.A., Brock, N. & West, K.D., 2007. Model Uncertainty and Policy Evaluation: Some Theory and Empirics. *Journal of Econometrics*, 136(2), pp.629-64.

Chmielecki, R.M. & Raftery, A.E., 2010. Probabilistic Visibility Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, 139(5), pp.1626-36.

Feldkircher, M., 2012. Forecast Combination and Bayesian Model Averaging: A Prior Sensitivity Analysis. *Journal of Forecasting*, 31(4), pp.361-76.

Fraley, C., Raftery, A.E. & Gneiting, T., 2010. Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging. *Monthly Weather Review*, 138(1), pp.190-202.

Gneiting, T. & Thorarinsdottir, T.L., 2010. *Predicting Inflation: Professional Experts Versus No-Change Forecasts*. Working Paper. http://arxiv.org/abs/1010.2318v1 (accessed June 15, 2011).

Koop, G. & Korobilis, D., 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).

Montgomery, J.M., Hollenbach, F.M. & Ward, M.D., 2012a. Improving Predictions Using Ensemble Bayesian Model Averaging. 20(3), pp.271-91.

Montgomery, J.M., Hollenbach, F.M. & Ward, M.D., 2012b. Say Yes to the Guess: Ensemble Methods to Predict Unemployment and Inflation. In *Proceedings of the 2012 Annual Meetings.* New Orleans, USA, Aug/Sept 2012b. American Political Science Association.

Montgomery, J.M. & Nyhan, B., 2010. Bayesian Model Averaging: Theoretical Developments and Practical Applications. *Political Analysis*, 18(2), pp.245-70.

Raftery, A.E., Gneiting, T., Balabdaoui, F. & Polakowski, M., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5), pp.1155-74.

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

Wright, J.H., 2008. Bayesian Model Averaging and Exchange Rate Forecasts. *Journal of Econometrics*, 146(2), pp.329-41.

Wright, J.H., 2009. Forecasting US Inflation by Bayesian Model Averaging. *Journal of Forecasting*, 28(2), pp.131-44.

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 Ph.D. candidate at Duke University 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. He first learned about forecasting from Kenneth Frank Janda.