

Bayesian Combination of State Polls and Election Forecasts*

Kari Lock¹ and Andrew Gelman²

¹ *Department of Statistics, Harvard University, lock@stat.harvard.edu*

² *Department of Statistics and Department of Political Science,
Columbia University, gelman@stat.columbia.edu*

21 September 2008

Abstract

In February of 2008, SurveyUSA polled 600 people in each state and asked who they would vote for in either head-to-head match-up: Obama vs. McCain, and Clinton vs. McCain. Here we integrate these polls with prior information; how each state voted in comparison to the national outcome in the 2004 election. We use Bayesian methods to merge prior and poll data, weighting each by its respective information. The variance for our poll data incorporates both sampling variability and variability due to time before the election, estimated using pre-election poll data from the 2000 and 2004 elections. The variance for our prior data is estimated using the results of the past nine presidential elections. The union of prior and poll data results in a posterior distribution predicting how each state will vote, in turn giving us posterior intervals for both the popular and electoral vote outcomes of the 2008 presidential election. Lastly, these posterior distributions are updated with the most recent poll data as of August, 2008.

Keywords: election prediction, pre-election polls, Bayesian updating, shrinkage estimation

1 INTRODUCTION

Research tells us that national elections are predictable from fundamentals (e.g., Rosenstone, 1983, Campbell, 1992, Gelman and King, 1993, Erikson and Wlezien, 2008), but this doesn't stop political scientists, let alone journalists, from obsessively tracking swings in the polls. The next level of sophistication—afforded us by the combination of ubiquitous telephone polling and internet dissemination of results—is to track the trends in state polls, a practice which was led in 2004 by Republican-leaning realclearpolitics.com and now in

*We thank the National Science Foundation, National Institutes of Health, and Columbia University Applied Statistics Center for partial support of this research.

2008 at fivethirtyeight.com, a website maintained by Democrat (and professional baseball statistician) Nate Silver.

Presidential elections are decided in swing states, and so it makes sense to look at state by state polls. On the other hand, the relative positions of the states are highly predictable from previous elections. So what is to be done? Is there a point of balance between the frenzy of daily or weekly polling on one hand, and the supine acceptance of forecasts on the other? The answer is Yes, a Bayesian analysis can do partial pooling between these extremes. We use historical election results by state and campaign-season polls from 2000 and 2004 to estimate the appropriate weighting to use when combining surveys and forecasts.

The year leading up to a presidential election is full of polls and speculation, necessitating a study of the measure of uncertainty surrounding predictions. Given the true proportion who intend to vote for a candidate, one can easily compute the variance in poll results based on the size of the sample. However, here we wish to compute the forecast uncertainty given the poll results of each state at some point before the election. To do this, we need not only the variance of a sample proportion, but an estimate for how much the true proportion varies in the months before the election, and a prior distribution for state-level voting patterns. We base our prior distribution on the 2004 election results and use these to improve our estimates and to serve as a measure of comparison for the predictive strength of pre-election polls.

In February of 2008, SurveyUSA (www.surveyusa.com) polled nearly 600 voters in each state, asking the questions “If there were an election for President of the United States today, and the only two names on the ballot were Republican John McCain and Democrat Hillary Clinton, who would you vote for?” and “What if it was John McCain against Democrat Barack Obama?” The poll was conducted over the phone using the voice of a professional announcer, with households randomly selected using random digit dialing (Survey Sampling International). Each response was classified as McCain, Clinton, Obama, or undecided. For each state the undecided category consisted of 5–14% of those polled, and these people as well as third-party supporters were excluded from our analysis. Likewise, for previous election results, we restrict the population to those who supported either the Democrat or the Republican.

This paper merges prior data (the 2004 election results) and the poll data described above. In sections 2 and 3 of this article, we ascertain the strength of each source of data in predicting the election. Section 2 focuses on our prior data, and contains an analysis of the use of past election results in predicting future election results, ultimately resulting in an estimate for the variance of the 2008 election results given the 2004 election results. Section 3 focuses on poll data, and contains an analysis of the strength of pre-election polls in predicting election results, giving measures both of poll variability and variability due to time before the election. Section 4 brings it all together with a full Bayesian analysis, fusing prior data with poll data to create a posterior distribution for the proportion voting for the Democrat in each state. In section 5 we update the posterior with more recent polling data and examine the implications for the 2008 election.

2 PAST ELECTION RESULTS

The political positions of the states are consistent in the short term from year to year; for example, New York has strongly favored the Democrats in recent decades, Utah has been consistently Republican, and Ohio has been in the middle. We begin our analysis by quantifying the ability to predict a state outcome in a future election using the results of past elections. We do this using the presidential elections of 1976–2004. We chose not to go back beyond 1976 since state results correlate strongly ($0.79 \leq r \leq 0.95$) for adjacent elections after 1972, while the correlation between the 1972 and 1976 elections is only 0.11.

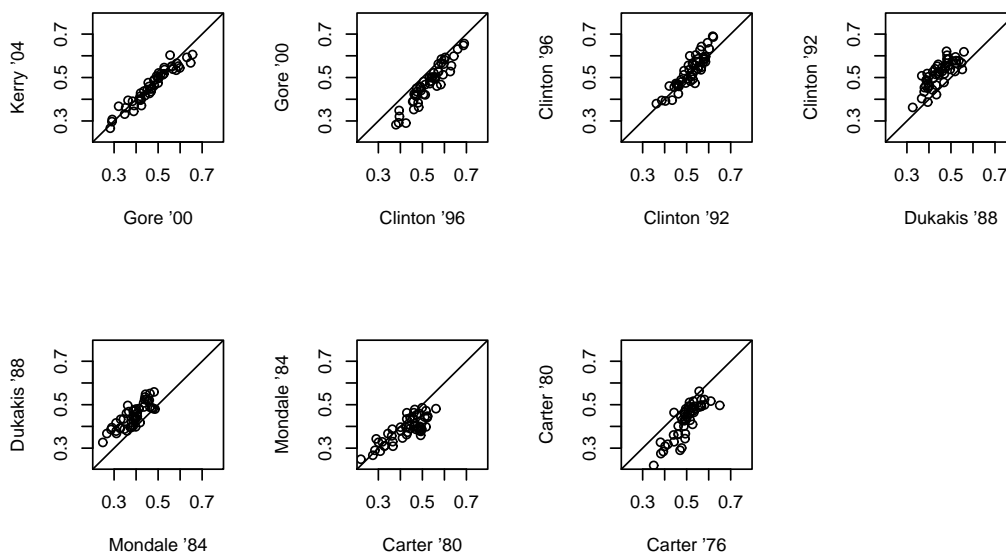


Figure 1: State results from one presidential election to the next, in each case showing the Democratic vote share in each state.

Figure 1 shows strong correlations in the Democratic candidate’s share of the vote in each state from one presidential election to the next. But in many cases the proportions are approximately uniformly shifted from the previous election. For example, states had much higher proportions for Clinton in 1992 than for Dukakis in 1988, and much lower proportions for Gore in 2000 than for Clinton in 1996. This does not indicate a change in state’s relative partisanship but rather a varying nationwide popularity of the Democratic candidate from election to election. The popularity of Kerry may not predict the popularity of Obama, but the popularity of Kerry in any given state compared to the popularity of Kerry nationwide seems to be indicative of the future popularity of Obama in that state as compared to nationwide. For this reason we look at the *state differences*, the difference between the proportion voting Democratic in each state and the national proportion voting Democratic.

We tried various exponential smoothing models utilizing a longer history of past elections to predict later elections, but the end result is that the most recent election alone is the best way to predict the next election. Therefore, in our analysis of 2008 we ignore election data before 2004, and simply consider the proportion of voters in each state choosing John Kerry over George W. Bush in the 2004 election. Our only adjustment is a home-state correction: we subtract 6% (as determined from past election results) of the vote for Bush and Kerry in Texas and Massachusetts, respectively, and give the same amount in the forecast for McCain in Arizona and Clinton in New York or Obama in Illinois. Finally, Kerry’s share of the two-party vote was 48.9% so our prior data become, for each state, the proportion voting for Kerry minus .489.

To determine the strength of our prior data, we need to know how much these state differences vary from election to election. Let $d_{s,y}$ be the state difference from the national proportion voting Democratic in state s , year y . We first estimate $var(d_{s,2008}|d_{s,2004})$ for each state by $\frac{1}{7} \sum_{i=1}^7 (d_{s,y_{i+1}} - d_{s,y_i})^2$, averaging over the seven presidential elections from 1976 through 2004. With only seven data points for each state, however, these estimates could be unreliable. We could get around this problem by assuming a common variance estimate for all states, but rather than forcing either one common estimate or fifty individual estimates, we use shrinkage estimation, partial pooling. Exactly how much to pull each estimate to the common mean is determined using a multilevel regression (using the `lmer` function developed by Douglas Bates in the statistical package R) and is based upon comparisons of within-state and between-state variability. Before pooling, the estimates of standard deviation for each state range from .011 to .073, with complete pooling the common estimate is .037; after our partial pooling the estimates range from .029 to .056.

From the normal approximation, we can expect the difference in 2008 to fall within .06 of the 2004 state difference for the most consistent states and up to .11 away for the least consistent states.

3 PRE-ELECTION POLLS

How much can we learn from a February poll of 600 voters in each state? If we ignore that the poll was conducted so early in the year, it appears we can learn quite a lot. Due to sampling variability alone, we would expect the true proportion who would vote Democrat in each state to be within .04 of the sample proportion ($SD = \sqrt{p(1-p)/n} \approx \sqrt{.5 * .5/600} = .02$). A standard deviation of .02 would make a poll of this size more informative than the 2004 election. Using Monte Carlo techniques, one could simulate many potential “true” proportions for each state, and so many potential popular or electoral college results, as done in Erikson and Sigman (2008). However, this would depict voter preferences *in February*. To get a true measure of variability, we need to consider not only sampling variability, but variability due to time before the election. Here we use pre-election polls from the months leading up to the 2000 and 2004 elections to get an estimate for this variance.

In both 2000 and 2004, the Annenberg Public Policy Center at the University of Pennsyl-

vanian conducted the NAES (National Annenberg Election Survey), a series of polls throughout the year leading up to the election. Again restricting our analysis only to those who say they would vote for the Democrat or the Republican, we have 43,373 people polled in 2000 and 52,825 in 2004.

Let $p_{s,t}$ represent the proportion of people in state s who intend, t months before the election, to vote for the Democrat. We estimate this with $\hat{p}_{s,t}$, based on the pre-election polls. Our focus here is estimating $var(p_{s,t}|p_{s,0})$, where $p_{s,0}$ is the proportion voting for Obama in state s at the time of the election. Since $p_{s,t}$ is unobserved we cannot estimate $var(p_{s,t}|p_{s,0})$ directly, so we empirically compute $var(\hat{p}_{s,t}|p_{s,0})$, and subtract off the expected sampling variability.

Formally,

$$\begin{aligned} var(\hat{p}_{s,t}|p_{s,0}) &= E[var(\hat{p}_{s,t}|p_{s,0}, p_{s,t})] + var[E(\hat{p}_{s,t}|p_{s,0}, p_{s,t})] \\ &= E[var(\hat{p}_{s,t}|p_{s,t})|p_{s,0}] + var[E(\hat{p}_{s,t}|p_{s,t})|p_{s,0}] \\ &= E\left[\frac{p_{s,t}(1-p_{s,t})}{n_{s,t}}\right] + var(p_{s,t}|p_{s,0}) \\ &= \frac{p_{s,0}(1-p_{s,0})}{n_{s,t}} + \left(\frac{n_{s,t}-1}{n_{s,t}}\right) var(p_{s,t}|p_{s,0}), \end{aligned}$$

so

$$var(p_{s,t}|p_{s,0}) = \frac{n_{s,t}}{n_{s,t}-1} \left(var(\hat{p}_{s,t}|p_{s,0}) - \frac{p_{s,0}(1-p_{s,0})}{n_{s,t}} \right). \quad (1)$$

If we had many polls each month in each state, we could estimate the right side of (1) empirically, and ideally we'd see a decreasing trend as the months get closer to the election. Unfortunately, we don't have that kind of data; for each state, each month, sample sizes range from 0 to 844, but with 42% having less than 30 people polled. In order to get a reliable estimate for variance, we need both larger sample sizes and multiple polls to average over.

We tried assuming $var(p_{s,t}|p_{s,0})$ to be constant for all s , allowing us to average over all states in a given month. Weighting by sample size, this gives the estimate

$$\widehat{var}(p_{s,t}|p_{s,0}) = \frac{\sum_{s=1}^{50} n_{s,t} * \frac{n_{s,t}}{n_{s,t}-1} \left((\hat{p}_{s,t} - p_{s,0})^2 - \frac{p_{s,0}(1-p_{s,0})}{n_{s,t}} \right)}{\sum_{s=1}^{50} n_{s,t}}. \quad (2)$$

Even when we pooled the states, however, sample sizes were too small to detect any kind of trend. Setting a variance decreasing approaching the election and using the sample sizes we have, simulation studies showed that the trend didn't start to shine through the noise until the standard deviation increased by at least .01 for each additional month prior to the election, while the actual increase seems to be closer to .002 or .003. Moreover, there is no guarantee that (2) will even be positive, and with such small sample sizes we were occasionally getting negative variance estimates.

To fix this, we decided to pool three months at a time, providing adequate sample sizes to ensure positive variance estimates and to illuminate the decreasing trend as common sense would predict. These estimates are displayed in Figure 2.

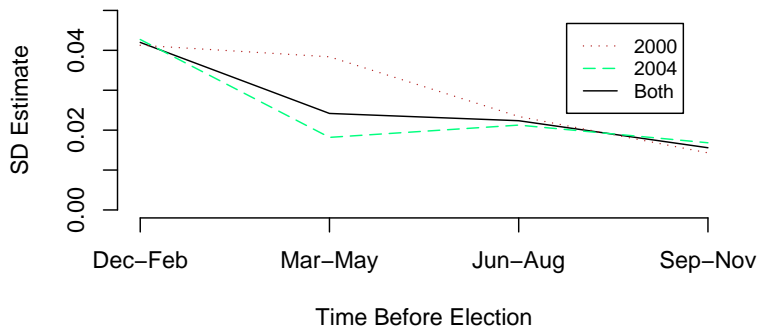


Figure 2: Estimated predictive standard deviation for the proportion of people who would vote for the Democratic candidate for president in a given state in the months leading up to an election, given the way the state voted in the actual election.

With the exception of March–May, the estimates agree surprisingly well for 2000 and 2004; whether this is coincidence or an indication of consistency in the true “variance due to time before the election” we cannot tell.

We tried to partially pool the estimates for each state, but this results in essentially complete pooling so we stick with the common estimate. For some states, these estimates get higher than the standard deviation estimated for the state outcome of the next election given the previous election (section 2). This may be due to giving all states the same estimate, or due to the imperfections of poll samples.

Fitting a least squares regression line to the estimates from both years combined, we get estimates of .038, .03, .022, and .014 for each of the four time periods respectively. Interpolating, our estimated standard deviation for February is .035.

4 POSTERIOR DISTRIBUTION

With the variance estimates derived in sections 2 and 3, we are all set to go forth with the full Bayesian analysis. Our February poll proportions follow a normal distribution (a reasonable approximation given the large sample size in each state), centered around the state proportion voting Democratic at the time of the election, and with variance given by (1), so

$$\hat{p}_{s,feb}|p_{s,0} \sim N\left(p_{s,0}, \frac{p_{s,0}(1-p_{s,0})}{n_{s,feb}} + .035\right). \quad (3)$$

The sample sizes range from 500 to 600, leading to standard deviations ranging from .055 to .057.

For the prior distribution, the 2004 election results tell us how we expect each state to vote compared to the popular vote, but not about the 2008 popular vote. For the 2008 popular vote prediction, we turn to fivethirtyeight.com, a website created by baseball analyst Nate Silver for predicting the outcome of the 2008 election. [Fivethirtyeight.com](http://fivethirtyeight.com) uses among other things a weighted combination of the major polls, weighted based on both date and reliability of the poll. As of August 13th, 2008, their predicted popular vote is .5107 for Obama. They give a margin of error for their predicted proportion of each state going Democratic, which we used with simulation to get a standard deviation for their popular vote estimate; .033. For Clinton, we use the predicted popular vote for her from our February polls, .506. The predicted popular vote for Obama based on the February polls is .514, but we use .5107 as the more recent estimate.

We combine the popular vote estimate with the estimated differences of each state from the popular vote to get the estimated proportion voting Democratic in each state. Thus our prior distribution is

$$p_{s,0}|d_{s,2004} \sim N(d_{s,2004} + .5107, \text{var}(d_{s,2008}|d_{s,2004}) + .033^2). \quad (4)$$

This gives a prior standard deviation ranging from .044 to .065. The sampling standard deviation of .056 lies in the middle of this range, and so the posterior is closer to the prior for some states and to the poll for other states, depending on the consistency of that particular state from election to election.

We use a normal-normal mixture model to create the posterior, weighting by information, the reciprocal of variance. Therefore our posterior is normal with mean

$$E(p_{s,0}|\hat{p}_{s,feb}) = \frac{\left(\frac{1}{\text{var}(\hat{p}_{s,feb}|p_{s,0})}\right)\hat{p}_{s,feb} + \left(\frac{1}{\text{var}(p_{s,0})}\right)(d_{s,2004} + \widehat{pop}_{2008})}{\frac{1}{\text{var}(\hat{p}_{s,feb}|p_{s,0})} + \frac{1}{\text{var}(p_{s,0})}}, \quad (5)$$

and variance

$$\text{var}(p_{s,0}|\hat{p}_{s,feb}) = \frac{1}{\frac{1}{\text{var}(\hat{p}_{s,feb}|p_{s,0})} + \frac{1}{\text{var}(p_{s,0})}}. \quad (6)$$

For a typical state, this simplifies to something like

$$p_{s,0}|\hat{p}_{s,feb} \sim N(.45\hat{p}_{s,feb} + .55(d_{s,2004} + .5107), .037^2).$$

with the weight on the polls ranging from .38 to .56. Figure 3 shows the posterior predictions (alongside prior and poll predictions) for both Clinton and Obama.

One of the great advantages of Bayesian analysis is the ability to incorporate new information. Besides the popular vote estimate in our prior, we've only used information available 10 months before the election. While the focus of this paper is not primarily predicting the election, but rather determining the appropriate weighting between past election results and poll data, given the ease of new data integration, it would be illogical to not utilize more recent data. Rather than sorting through countless polls conducted between

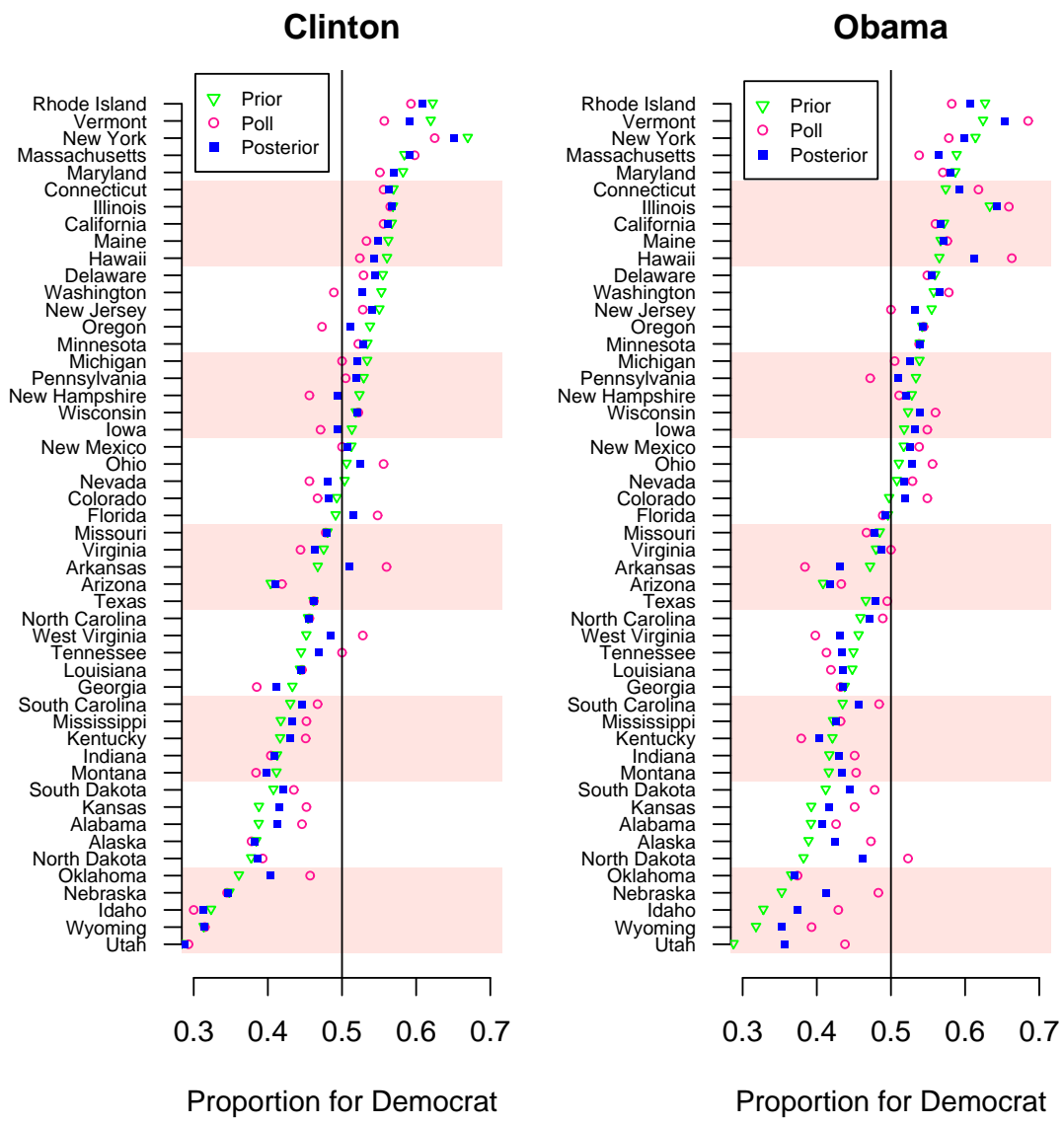


Figure 3: Prior, poll, and posterior predictions for the proportion voting Democratic in each state, for both Barack Obama and Hillary Clinton. States are ordered by decreasing proportion voting Democratic in the 2004 election.

February and August, we rely on the hard work of fivethirtyeight.com doing precisely that. Our old posterior now becomes our new prior, and the estimates (and variability) of fivethirtyeight.com become our new data, representing a more current prediction. The weighting is between 46% and 59% on our old posterior, with the rest on the fivethirtyeight.com estimates. These updated posterior intervals are displayed in Figure 4(a), above our original posterior intervals for each state.

As seen in Figure 4(a), our point predictions don't change much in our posterior updating (in general, the fivethirtyeight.com estimates were already quite close to our original posterior estimates). The variability in our estimates is less in our updated posterior, due to the added information. Interestingly, in determining the probability each state goes Democratic, more important than the predicted proportion voting Democratic in each state is the measure of variability on that prediction. As seen in Figure 4(b), the probability of going Democratic is usually higher in our updated posterior in states leaning Democratic, and lower in states leaning Republican, since a smaller variance makes a state more likely to go whichever way it is leaning.

5 PREDICTING THE ELECTION

Now that we have posterior predictions for each state, what does this imply for the outcome of the election? What matters in terms of the election outcome is not the proportion voting Democratic in each state, but whether or not this proportion is greater than .5.

We run 100,000 simulated elections, giving posterior distributions for the popular and electoral votes. These distributions are displayed in Figure 5 and summarized in Table 1 alongside the simulation results based on each of our data sources.

CLINTON	Prior	Poll	Posterior		
Popular vote	.509 (.490, .527)	.506 (.484, .528)	.508 (.494, .522)		
Electoral vote	.529 (.396, .651)	.520 (.370, .651)	.533 (.413, .643)		
Pr(Clinton Win)	.693	.639	.739		
OBAMA	Prior	Poll	Old Post	538.com	New Post
Popular vote	.512 (.493, .531)	.514 (.492, .536)	.513 (.499, .528)	.512 (.497, .527)	.513 (.502, .523)
Electoral vote	.549 (.418, .669)	.554 (.400, .690)	.558 (.441, .673)	.554 (.446, .662)	.556 (.467, .651)
Pr(Obama Win)	.795	.782	.844	.846	.893

Table 1: Results based on simulated elections of Clinton vs. McCain and Obama vs. McCain polls. Estimates are for the general election based on information before the Democratic nomination had been resolved. For popular and electoral votes the top numbers are means and the parentheses give 95% posterior intervals.

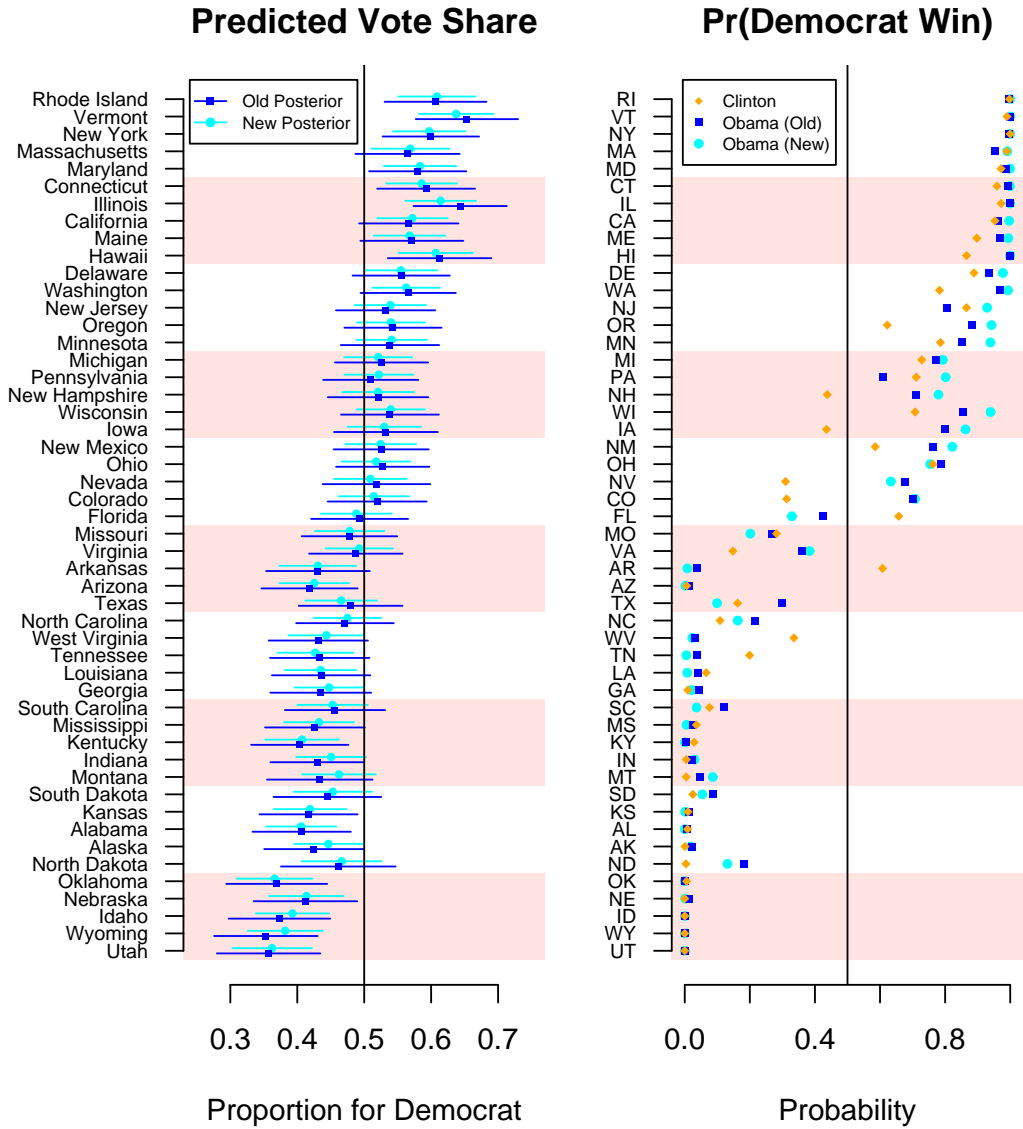


Figure 4: (a) Estimates and 95% posterior intervals, both original and updated, for the proportion voting for Obama in each state. (b) The estimated probability of each state going for Obama. States are ordered by decreasing proportion voting Democratic in the 2004 election.

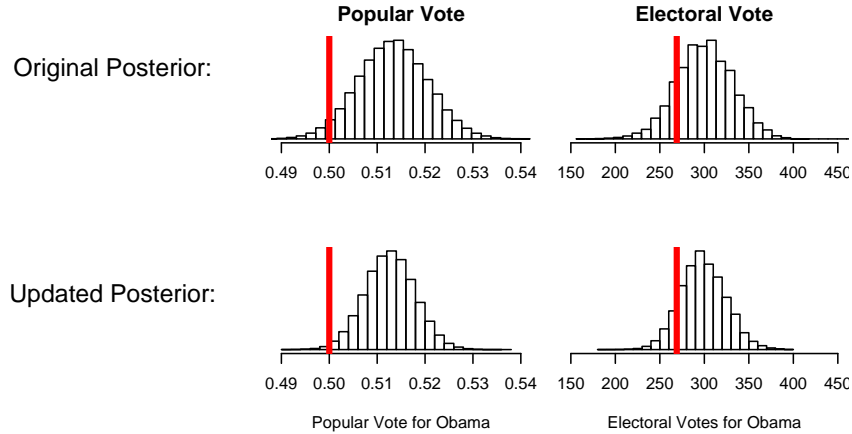


Figure 5: Posterior distributions for the popular and electoral vote. Electoral votes to the right of the solid line represent an Obama victory.

The overall probability that Obama wins is higher as we get more certain in our predictions for two reasons; one being that the entire election prediction favors the Democrats, and the second being that more of the swing states lean Democratic.

6 Conclusion

This paper has the goal of determining the strength of past elections and of pre-election polls in predicting a future election, and combining these sources to predict the election.

We found that, to predict the current election, using only the most recent election is more effective than a weighted average of previous election results, and that this is a good predictor of the way each state votes compared to the nation, but not necessarily of the national vote. Hence, past election data is best used with a current estimate of the popular vote (such as can be obtained from polls or from forecasts that use economic and other information). Thus, our key contribution here is to separate the national forecast (on which much effort has been expended by many researchers) from the relative positions of the states (for which past elections and current polls can be combined in order to make inferences).

Pre-election polls, not surprisingly, are more reliable as they get closer to the election. In general, the weighting between past election data and pre-election polls depends on many things; the consistency of the state from election to election, the sample size of the poll, the month the poll was conducted, and the reliability of the popular vote estimate.

References

- [1] Campbell, J. E. (1992). “Forecasting the Presidential Vote in the States.” *American Journal of Political Science* 36: 386-407.

- [2] Erikson, R. S., and Sigman, K. (2008). “Guest Pollster: The SurveyUSA 50 State Poll and the Electoral College.” Pollster.com, http://www.pollster.com/blogs/guest-pollster_the-surveyusa_5.php, 3/08.
- [3] Gelman, A., and King, G. (1993). “Why are American Presidential Election Campaign Polls so Variable When Votes are so Predictable?” *British Journal of Political Science* **23**: 409-451.
- [4] Rosenstone, S. J. (1983). *Forecasting Presidential Elections*. Yale University Press.
- [5] Silver, N. (2008). <http://www.fivethirtyeight.com/>, 8/08.
- [6] Wlezien, C., and Erikson, R. S. (2007). “The Horse Race: What Polls Reveal as the Election Campaign Unfolds.” *International Journal of Public Opinion Research* 19: 74-88.
- [7] Wlezien, C., and Erikson, R. S. (2004). “The Fundamentals, the Polls, and the Presidential Vote.” *PS: Political Science and Politics* **37**: 747-751.
- [8] www.surveysampling.com, 6/08.
- [9] <http://www.annenbergpublicpolicycenter.org/AreaDetails.aspx?myId=1>, 6/08.