ARTICLE IN PRESS

International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Origins of Presidential poll aggregation: A perspective from 2004 to 2012

Samuel S.-H. Wang *

Princeton Neuroscience Institute and Department of Molecular Biology, Princeton University, Princeton, NJ 08544, United States

ARTICLE INFO

Keywords:
Poll aggregation
Bayesian estimation
Opinion polls
Election forecasting
Election snapshots

ABSTRACT

US political reporting has become extraordinarily rich in polling data. However, this increase in information availability has not been matched by an improvement in the accuracy of poll-based news stories, which usually examine a single survey at a time, rather than providing an aggregated, more accurate view. In 2004, I developed a meta-analysis that reduced the polling noise for the Presidential race by reducing all available state polls to a snapshot at a single time, known as the Electoral Vote estimator. Assuming that Presidential pollsters are accurate in the aggregate, the snapshot has an accuracy equivalent to less than ±0.5% in the national popular-vote margin. The estimator outperforms both the aggregator FiveThirtyEight and the betting market InTrade. Complex models, which adjust individual polls and employ pre-campaign "fundamental" variables, improve the accuracy in individual states but provide little or no advantage in overall performance, while at the same time reducing transparency. A polls-only snapshot can also identify shifts in the race, with a time resolution of a single day, thus assisting in the identification of discrete events that influence a race. Finally, starting at around Memorial Day, variations in the polling snapshot over time are sufficient to enable the production of a high-quality, random-drift-based prediction without a need for the fundamentals that are traditionally used by political science models. In summary, the use of polls by themselves can capture the detailed dynamics of Presidential races and make predictions. Taken together, these qualities make the meta-analysis a sensitive indicator of the ups and downs of a national campaign—in short, a precise electoral thermometer.

1. Introduction

In 2012, polling aggregation entered the public spotlight as never before. Typically, political horserace commentaries in the US are dominated by pundits who are motivated by pressure, not to be accurate, but to attract readers and viewers. For example, one day before the 2012 U.S. presidential election, former Reagan speechwriter Noonan (2012) wrote that "nobody knows anything" about who would win, asserting that Republican candidate Mitt Romney's supporters had the greater passion and enthusiasm, while columnist George Will predicted a Romney electoral landslide (Poor, 2012).

In the end, the aggregators were correct. The pundits largely failed to report the fact that, according to public opinion polls with collectively excellent track records, President Obama had an advantage of three to four percentage points for nearly the entire campaign season. Ignoring the data, many commentators expressed confidence—and were wrong.

* Correspondence to: Princeton Neuroscience Institute, Washington Road, Princeton University, Princeton, NJ 08544, United States. Tel.: +1 609 258 0388; fax: +1 609 258 1028.
E-mail address: sswang@princeton.edu.

http://dx.doi.org/10.1016/j.ijforecast.2015.01.003
0169-2070/

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

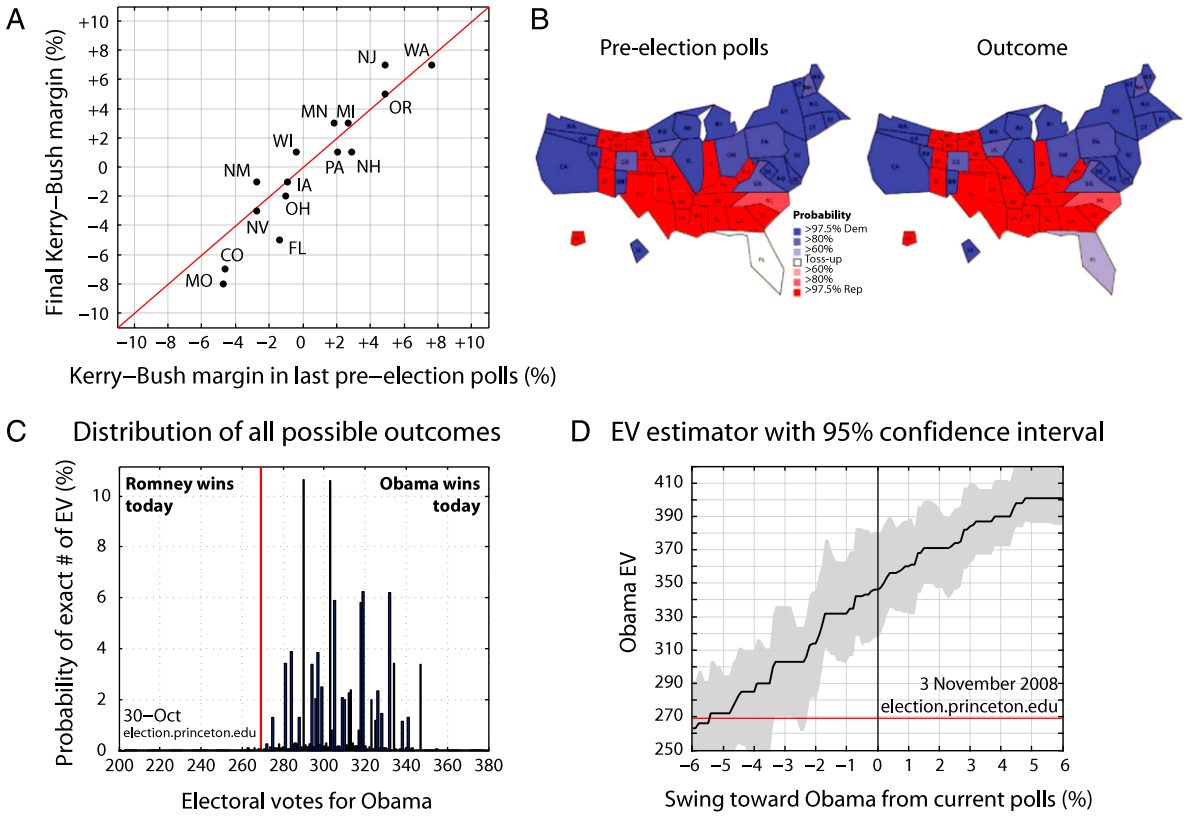

Fig. 1. Foundations of the Presidential meta-analysis. (a) State-by-state election margins as a function of final pre-election polls in the 2004 Kerry vs. Bush race. (b) Pre-election win probabilities and actual outcomes in the 2012 Obama vs. Romney race. (c) A snapshot of the exact distribution of all $2^{51} = 2.3$ quadrillion outcomes calculated from the win probabilities in (b). The electoral vote estimator is defined as the median of the distribution. (d) Electoral effect of a uniform shift in state polls through a constant swing. The gray band indicates a nominal 95% confidence interval, including uncorrected pollster-to-pollster variation.

In this article, I describe an early approach to the aggregation of Presidential state polls, the meta-analytic method, which has been being used at the Princeton Election Consortium (PEC; http://election.princeton.edu) since 2004. PEC's approach uses Electoral College mechanisms and can be updated on a daily basis. Its only input is publicly available data, and it runs on open-source software, thus providing a high level of transparency. I will describe this method, and give both public and academic perspectives (see also Jones, 2008, for a review). I provide both an academic account and a history, under the assumption that the evolution of the meta-analysis may interest some readers.

Polling aggregators have been outperforming pundits since at least 2004, when a number of websites began to collect and report polls on a state-by-state basis in Presidential, Senate, and House races. State polls are of particular interest for the Presidency, for three reasons. First, the Presidency is determined via the Electoral College, which is driven by state election win-lose outcomes. Second, state polls have the advantage of being accurate predictors of state election outcomes, on average (Fig. 1(a)), though national polls can have significant inaccuracies. For example, in 2000, Al Gore won the popular vote over George W. Bush by 0.5%, yet election-eve national polls favored Bush by an average of 2.5%, a 3.0% error that got the sign of the outcome wrong. State polls may owe their superior accuracy levels to the fact that local populations are less complex demographically, and therefore easier to sample, than the nation as a whole. Third and last, state presidential polls are also remarkably abundant: Electoral-vote.com contains the results of 879 polls from 2004, 1189 from 2008, and 924 from 2012.

Early sites—RealClearPolitics in 2002, followed in 2004 by Andrew Tanenbaum's Electoral-vote.com, the Princeton Election Consortium, and several others (Forelle, 2004a)—reported average or median polling margins (i.e., the percentage difference in support between the two leading candidates) for individual races. An additional step was taken by PEC (then titled "Electoral college meta-analysis", http://synapse.princeton.edu/~sam/pollcalc.html), which calculated the electoral vote (EV) distribution of all possible outcomes, using polls to provide a simple tracking index, the EV estimator. The calculation, an estimate of the EV outcome for the Kerry vs. Bush race, was updated in a low-graphics, hand-coded HTML webpage, together with a publicly posted MATLAB script. PEC gained a following among natural scientists, political and social scientists, and financial analysts. Over the course of the 2004 campaign, PEC attracted over a million visits, and the median decided-voter calculation on election eve captured the exact final outcome (Forelle, 2004b).

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

In 2008, a full PEC website, unveiled under the banner "A first draft of electoral history", provided results based on decided-voter polling from all 50 states, as well as Senate and House total-seat projections. In the closing week of the campaign, PEC ended up within one electoral vote of the final outcome, within one seat in the Senate, and exactly correct in the House.

The same year, many other aggregators emerged on the scene. The website 3BlueDudes.com documented at least 45 different hobbyists in 2008. One site rapidly emerged as the most popular: FiveThirtyEight. Created by sabermetrician Nate Silver, FiveThirtyEight arose from his efforts on the liberal weblog DailyKos. Silver initially attracted attention for his analysis of the Democratic nomination contest between Hillary Clinton and Barack Obama. In the general election season, Silver provided a continuous feed of news and lively commentary, as well as a prediction of the November outcome based on a mix of economic, political, and demographic assumptions ("fundamentals"), along with the polling data. FiveThirtyEight was later licensed to the *New York Times* from 2010 to 2012, becoming a major driver of traffic to the *Times* website (Tracy, 2012).

In the academic sector, fundamentals and polling data have long been used to study Presidential campaigns. Most academic research has focused on time scales of months or longer, usually concentrating on explaining outcomes after the election, or on making predictions before the start of the general election campaign. Predictions are usually done in the spirit of testing provisional models which then are subject to change (for reviews, see Abramowitz, 2008; Jones, 2008; Lewis-Beck & Tien, 2008; and articles in the current issue of the *International Journal of Forecasting*). In short, such models ask why elections turn out as they do.

However, for purposes of tracking and everyday prediction, such models suffer from several deficiencies. First, they have a lower time resolution than even a month-to-month pace, and are designed to be used once per election year, before the campaign starts. Second, they typically only make national-level predictions, and are based on very small numbers of past observations, i.e., however many Presidential elections have taken place in the baseline period. This may limit their confidence and accuracy. Indeed, an aggregate of fundamentals-based models in October 2012 could only predict President Obama's 2012 reelection with a 60% probability (Montgomery, Hollenbach, & Ward, 2012), whereas the meta-analysis had been giving probabilities of above 90% since the summer of that year.

Polls-only analyses have been performed by Gelman and King (1993), who analyzed time trends from national polling data. Since 1996, Erikson and Wlezien (2012) have constructed detailed time series for the production of post-hoc trajectories of national campaigns. Using Electoral College mechanisms and state polls, Soumbatiants (Soumbatiants, 2003; Soumbatiants, Chappell, & Johnson, 2006) calculated a distribution of probable EV outcomes using Monte Carlo simulations, and examined the effects of hypothetical single-state or nationwide shifts in opinion. These scenarios have also been explored from the point of view of a campaign (Strömberg, 2002) or of an individual voter (Gelman, Silver, & Edlin, 2010). Strömberg (2002) correctly noted the pivotal nature of Florida in the final outcome, and found that campaigns allocated resources in a manner that scaled with the influence of individual states.

In 2012, day-to-day forecasting took three forms. First, the Princeton Election Consortium took a polls-only approach. Drew Linzer (http://votamatic.org; Linzer, 2013) took a second approach, using pre-election variables to establish a prior win probability and updating this in a Bayesian manner using new polling data. The resulting prediction was notably stable for the entire season. Extensive modeling was also done by Simon Jackman and Mark Blumenthal at the Huffington Post (Jackman & Blumenthal, 2013). In the public sphere, FiveThirtyEight combined prior and current information in order to create a measure that contained mixed elements of both snapshot and fundamentals-based prediction in a single measure. As of 2014, these organizations and others continue to analyze polls (Altman, 2014).

2. Data

The PEC core calculation is based on publicly available Presidential state polls, which are used to estimate the probability of a Democratic/Republican win on any given date. These are then used to calculate the probability distribution of the electoral votes corresponding to all $2^{51} = 2.3$ quadrillion possible state-level combinations.

Data sources and scripts. The polling data came from manual curation (2004), an XML feed from Pollster.com (2008), and a JSON feed from Huffington Post/Pollster (2012). In all cases, the data source was selected so as to include as many polling organizations as possible, with no exclusions. When both likely-voter and registered-voter values were available for the same poll, the likely-voter result was used. For the District of Columbia, no polls were available and the win probability for the Democratic candidate was assumed to be 100%. All scripts for data analysis and graphics generation were written in MATLAB and Python, and have been posted at http://election.princeton.edu, and deposited at the github software archive. In 2004, updates were done manually. In 2008 and 2012, updates were done automatically from July to election day. The update frequency increased as election day approached, with up to six updates per day in October.

3. Method

3.1. An exact calculation of the probability distribution

The win probability for any given state s at time t is termed $p_s(t)$, and is assumed to be predicted by the polling margin. Polling margins and analytical results are reported, using the sign convention that a positive number indicates a Democratic advantage. For any given date, p_s was determined using either the three most recent polls, or one week's worth of polls, whichever was greater. A three-poll minimum was chosen to reflect the fact that only closely-contested states had more than a few polls per month, and not until October even in those cases. The one-week criterion represents a tradeoff between capturing

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

enough polls to minimise the uncertainty and allowing movements in opinion to be detected quickly; one week also represents the length of a single news cycle. A poll's date was defined as the middle date on which it took place; if the oldest two had the same date, four polls were used. The same pollster could be used more than once for a given state if the samples contained non-overlapping respondent populations.

From these inputs, a median margin (M_s) was calculated. The median was used instead of the mean in order to prevent outlier data points from erroneous or methodologically unsound individual polls from having undue influence. More broadly, the use of the median takes the place of estimating and correcting pollster biases, an approach that is somewhat opaque and does not solve the issue of what to do with polling organizations that produce only one or a few polls. The estimated standard error of the median (σ_s) was calculated as $SD_s = 1.485 * (\text{median absolute deviation})/\sqrt{N}$. The Z-score, M_s/σ_s, was converted to a win probability p_s (Fig. 1(b)) using the t-distribution. (Note that the original calculations published in 2004 used the normal distribution. However, all calculations in this manuscript, including those for 2004, use the t-distribution.)

The probability distribution of all possible outcomes, $P(EV)$ (Fig. 1(c)), was calculated using the coefficients of the polynomial

$$\prod_s ((1 - p_s) + p_s x^{E_s}), \tag{1}$$

where $s = 1 \ldots 51$ represents the 50 states and the District of Columbia, and E_s is the number of electoral votes for state s. In this notation, x is a placeholder variable, such that the coefficient of the x^N term is the probability of winning a total of N electoral votes, $P(EV = N)$. The fact that electoral votes are assigned on a district-by-district basis in Nebraska and Maine was not taken into consideration. The median of P was used as the EV estimator.

The same approach was taken for modeling Senate outcomes, with $E_s = 1$ for all races. In addition, for modeling House outcomes, races were scored as $p = 0.5$ for toss-ups as defined by Pollster.com and set to $p = 0$ or $p = 1$ otherwise, giving a 68% confidence interval of $\pm\sqrt{N}$ seats for N toss-up races.

3.2. A polling bias parameter and the popular meta-margin

The snapshot win probability, defined as the probability of one candidate getting 270 or more electoral votes out of 538, was usually over 99% for one candidate or the other on a given day. A quantity that varied more continuously, and was therefore more informative, was the popular meta-margin (MM). MM is defined as the amount of opinion swing, spread equally across all polls, that would bring the median electoral vote estimator to a 269–269 tie. To identify the tie point, $P(EV)$ was calculated by varying the offset x over a range, i.e., by replacing M_s with $M_s + x$ (Fig. 1(d)).

It should be noted that, because voter demographics and perceptions vary from state to state, real shifts in opinion are not distributed evenly across all states. Thus, the meta-margin only approximates the magnitude of the true national shifts. Nonetheless, it has useful applications. The meta-margin allows the analysis of possible biases in polls. For example, if polls understate the support for one candidate by 1%, this would reverse the prediction if the meta-margin were less than 1% for the other candidate. As a second example, if 1% of voters switch from one candidate to the other, this generates a swing of 2% and can compensate for a meta-margin of 2%. In this way, the popular meta-margin is equivalent to the two-candidate difference found in single polls, allows evaluation of a wide variety of polling errors, and provides a mechanism for making predictions.

3.3. Prediction of November outcomes

The prediction for 2012 was produced under the assumption that the random drift followed historical patterns for Presidential re-election races. The amount of change between the various analysis dates between June 1 and election day was modeled using a bias parameter b applied across all polls, i.e. using margins of $M_s + b$ instead of M_s. Therefore, the win probability is the probability that $MM - b > 0$.

b was assumed to follow a t-distribution, setting the number of degrees of freedom equal to three. The t-distribution has longer tails than the normal distribution, and was chosen in order to incorporate mathematically the possibility of outlier events such as the 1980 Reagan–Carter race, during which the standard deviation of the two-candidate margin was ∼6% based on national polls (Erikson & Wlezien, 2012). The 2012 distribution of b was estimated using the 2004 meta-analysis, as a re-election year in which the meta-margin had a standard deviation (MMSD) of 2.2%. In historical data based on national polls, a similar stability can be found in pooled trajectories of re-election races from multiple pollsters (see Figure 2.1 of Erikson & Wlezien, 2012). However, it is difficult to estimate MMSD from national data, due to sampling error. For example, Gallup national data showed a standard deviation of 4.9% in 2004, and standard deviations of between 2.9 and 4.9% in six re-election races from 1972 to 2004.

3.4. Voter power

The power of a single voter in state s was determined by calculating the incremental change in one candidate's election-win probability $\Delta P_s(EV \geq 270)$ arising from a change in M_s of a fraction of a percentage point, and normalized by the number of votes cast in the most recent Presidential election. ΔP_s for each state was normalized to voters in the most powerful state or to one voter in New Jersey. The latter measure was termed a "jerseyvote".

3.5. Tracking national opinion swings

In order to track national opinion swings with a high time resolution (Fig. 8), all national polls within a given time interval were divided equally into single-day components, then averaged for each day without weighting, to generate a time course. After the election, the time course was adjusted by a constant amount to match the actual popular-vote result.

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

4. Results

4.1. Kerry vs. Bush 2004: an initial estimate of the bias variable

The first version of the meta-analysis, published starting in July 2004, analyzed the closely-fought re-election race of President George W. Bush (R) against his challenger, Senator John Kerry (D). The meta-analysis was announced on DailyKos.com and almost immediately attracted thousands of readers, and for good reason. The race was close and suspenseful, and the EV estimator crossed the 270 EV threshold three times during the general election campaign (Fig. 2(a)). The meta-analysis was necessary to enable this to be seen, since the swings were not large in terms of popular support: a one-point change in the two-candidate margin across all states caused a change of 30 EV in the electoral margin. On election eve, the polls-only estimate (i.e., an estimate with bias parameter $b = 0\%$) turned out to be exactly correct: Bush 286 EV, Kerry 252 EV. Because the smallest single-state margin was 0.4% (Wisconsin), the uncorrected meta-analysis had an effective accuracy of less than $\pm 0.4\%$ in units of popular opinion.

During the campaign, sharp or substantial moves in the EV estimator occurred after the Democratic convention (but not the Republican convention), the Swift Boat Veterans for Truth advertising campaign, and the first Presidential debate. The later debates had little effect, and the race was static from October 7th onward.

Despite the accuracy of the polls-only meta-analysis, I personally made an erroneous prediction. In the closing weeks of the campaign, I suggested that undecided voters would vote against the incumbent, a tendency that had been noticed in earlier campaigns. This led me to make an estimate of $b = +1.5\%$ toward John Kerry, which led to an incorrect prediction of Kerry 283 EV, Bush 255 EV. The incumbent rule, which was derived from an era in which recent pre-election polls were often not available, was therefore rejected for subsequent analyses. I also concluded that interpreting polling data is susceptible to motivated reasoning and biased assimilation, cognitive biases that occur even among quantitatively sophisticated persons (Kahan, Peters, Dawson, & Slovic, 2013). These reasons lead me to strongly recommend setting b to zero for tracking purposes.

The bias variable b was still useful for readers who wanted to apply alternative scenarios. If a reader thought that turnout efforts would boost his/her candidate by N points, that could be added as $b = N$ and the script recalculated. If he/she thought that one candidate would gain N points at the expense of the other, b could be set equal to $2N$. A map on the PEC website showed the effect of $b = \pm 2\%$. For other scenarios, more sophisticated readers could download and modify the MATLAB code.

4.2. Alternative scenarios and the jerseyvote index

As formulated in Eq. (1), exploring alternative scenarios is easy. The most straightforward approach is to alter p_s directly by setting its value to 0 ("what if Romney wins Florida?") or 1 ("what if Obama wins Florida?"). Alternately, the polling margin M_s can be shifted for one or more states.

In 2004, this perturbation approach was introduced via the concept of the "jerseyvote", a fanciful way of expressing the concept of individual voter power. The jerseyvotes calculation was done by shifting all polls to create a near-tied race, adding an additional small change in M_s in a single state, and calculating the resulting change in the win probability. Conceptually, jerseyvotes are related to the Penrose-Banzhaf power index (Banzhaf, 1965). Jerseyvotes express an individual's relative power to influence the final electoral outcome. For example, if a voter in Colorado has ten times as much influence over the national win probability as a voter in New Jersey, the Coloradan's vote is worth 10 jerseyvotes. Sadly for the hosts of PEC, one jerseyvote is not worth very much. PEC advised New Jersey residents to vote early, then amplify their efforts by tens of thousands by helping Pennsylvania voters get to the polls. In 2008 and 2012, readers were provided with a voter influence table (Table 2).

4.3. Accuracy in off-year elections, 2006 and 2010

Based on 2004 and 2008, state polls are highly accurate in the aggregate. However, are they accurate in off-year elections as well? In 2006, using simple polling medians and a compound probability calculation, I estimated the probability of a Democratic takeover of the US Senate at approximately 50%, a higher chance than that predicted by either pundits or electronic markets. The Democrats (along with two independents) took control of the Senate, with a 51–49 majority. I did not make a House prediction.

In the 2010 midterm off-year election, all of the Senate races were called correctly, with the exception of the re-election race of Senator Harry Reid (D-NV) against Sharron Angle, in which Reid trailed in the last eight pre-election polls, yet won by over five points. This polling error has been ascribed to an under-sampling of cell-phone-only and Hispanic voters.

In 2014, PEC's 2014 Election Eve Senate snapshots erred in the opposite direction as 2010, underestimating Republican-over-Democratic margins in multiple races, and showing Greg Orman (I-KS) and Kay Hagan (D-NC) in the lead. These results indicate that state-level surveys lose accuracy in off-year elections.

In the 2010 House election, Republicans retook control, with a 51-seat margin. PEC used district-by-district pre-election polls to predict a 25-seat Republican margin, a substantial underestimate. Most analysts performed similarly, suggesting that district-specific polls may not capture differences in voter intensity between the parties in an off-year. Congressional generic preference polls on election eve showed an average seven-point advantage for Republicans, which would have led to a more accurate prediction.

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

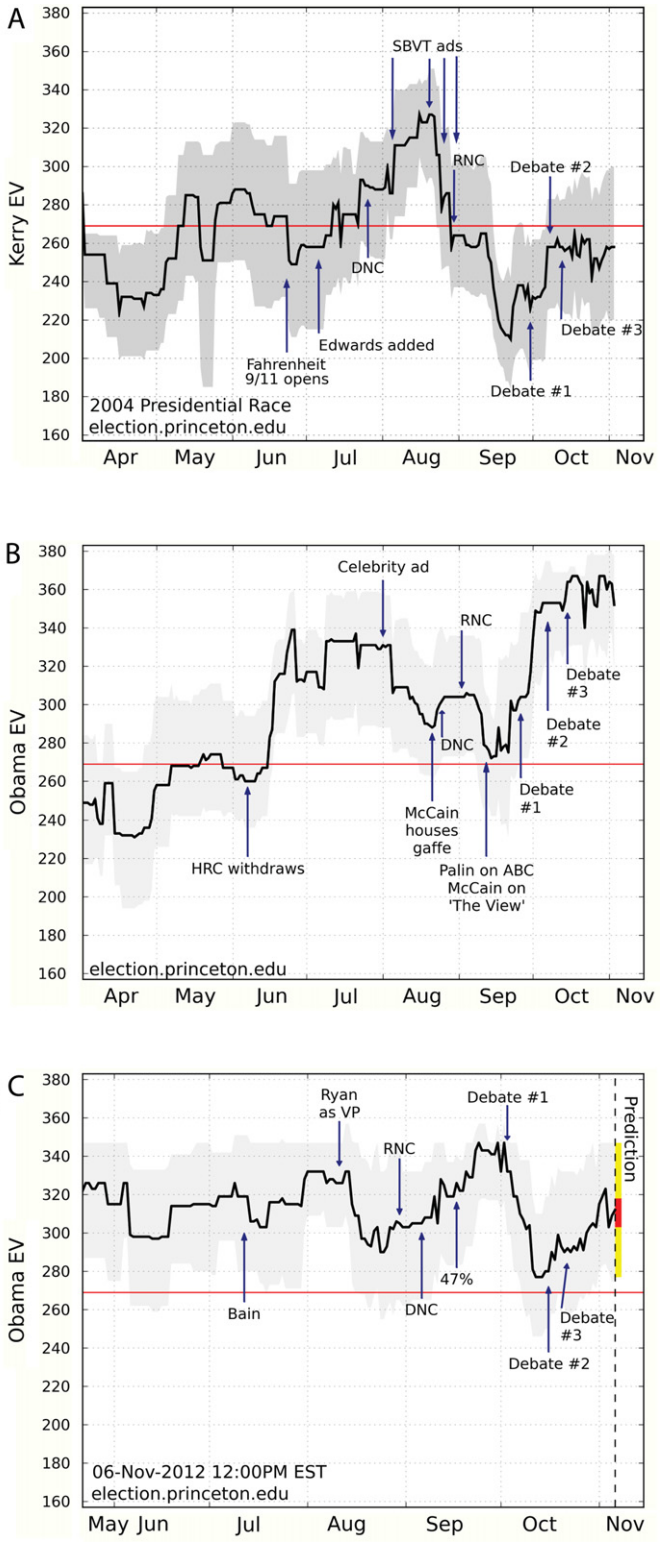

Fig. 2. Time series of the meta-analytic electoral vote predictor, 2004–2012. The EV estimator for the most recently available state polls, plotted as a function of time for (a) 2004, (b) 2008, and (c) 2012. The arrows indicate notable campaign events. Upward-pointing arrows indicate events that are likely to benefit the Democratic candidate, downward-pointing arrows the Republican candidate. DNC = Democratic National Convention; RNC = Republican National Convention; SBVT = Swift Boat Veterans for Truth ad campaign; HRC = Hillary Rodham Clinton. The gray band indicates a nominal 95% confidence interval, including uncorrected pollster-to-pollster variation.

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

Table 1
Comparison of the polling meta-analysis with election outcomes, 2004–2012. The win probability was calculated under the assumption of a symmetric drift (t-distribution, three degrees of freedom) with $\sigma = 2.2\%$ between July 1 and election day. The meta-margin standard deviation was calculated from June 1 to election day. National polls were calculated as the median of all polls conducted between November 1 and election day.

Year	PEC forecast/snapshot			National polls	Outcome	
	July 1 Democratic win probability	November 1 Democratic EV estimate	November 1 Meta-margin MM (SD)	Poll median	Democratic EV	Popular vote outcome (two-party)
2000				Bush +2.5%	266 EV	Gore +0.5%
2004	38%	252 EV	Bush +0.7% (1.2%)	Bush +2.0%	252 EV	Bush +3.0%
2008	90%	364 EV	Obama +8.0% (2.2%)	Obama +7.5%	365 EV	Obama +7.3%
2012	90%	315 EV	Obama +2.6% (1.2%)	Tie (+0.0%)	332 EV	Obama +4.0%

Table 2
The power of an individual voter. As an example calculation, a listing of voter power as calculated on election eve, November 5, 2012.

State	Median polling margin	Power
NH	Obama +2%	100.0
IA	Obama +2%	82.2
PA	Obama +3%	77.8
OH	Obama +3%	74.0
NV	Obama +5%	71.9
VA	Obama +2%	71.0
CO	Obama +2%	63.7
WI	Obama +4.5%	44.7
NM	Obama +6%	30.1
FL	Tied	26.6
MI	Obama +5.5%	21.6
OR	Obama +6%	19.1
NC	Romney +2%	5.2
MN	Obama +7.5%	3.2
LA	Romney +13%	0.9
NJ	Obama +12%	0.00091

4.4. Obama vs. McCain 2008: identifying a campaign's turning points

The algorithm was kept the same in 2008, except for the addition of automatic updates to enable the graphical tracking of time trends on a daily basis. This calculation used polling data for all 50 states and the District of Columbia, resulting in the electoral histories shown in Fig. 2.

Both the EV estimator and the meta-margin showed Senator Barack Obama (D) to be ahead for almost the entire general election campaign, with an electoral lead of 20 to 200 electoral votes and one to eight percentage points. At times, though, this lead shifted rapidly (Figs. 2(b), 8). Senator McCain (R) immediately gained a large but transient benefit from the addition of Alaska Governor Sarah Palin as his running mate. Following her riveting convention speech, the meta-analysis moved from a large Obama lead to a near-tie. Considering the delays in getting fresh state-level data, it is possible that McCain led Obama at this time. However, the EV estimator reversed its course shortly after Palin's unsuccessful interview with Charlie Gibson on ABC. After that, the movement toward Obama accelerated after the collapse of Lehman Brothers, a defining event of that year's economic crash. This movement toward Obama continued after the first Presidential debate, and for the rest of October.

By election day, the EV estimator had stabilized at 353 EV for Obama, with a nominal 68% confidence band of [337, 367] EV and a 95% confidence band of [316, 378] EV. These confidence bands included pollster variation (house effects), and so the true uncertainty was likely to be substantially lower. Using a wider time window in order to minimize the variance in the time series gave a final estimate of 364 EV (Table 1), just one electoral vote away from the final outcome, Obama 365 EV, McCain 173 EV. The final meta-margin, Obama +8.0%, was close to the final national polling median, indicating Obama +7.5%. Obama's final margin in the national popular vote was +7.3%.

Downticket, the polls showed comparable overall levels of accuracy (Table 3). In the Senate, the median outcome was 58–59 Democratic+Independent seats, with the Minnesota race (D-Franken vs. R-Coleman) being too close to call. The final outcome was 59 Democratic+Independent seats. In the House, taking all polls available at Pollster.com and assigning each winner to the leader, the Democrats were predicted to win 257 ± 3 seats (68% confidence interval, 254–260 seats) assuming binomial random outcomes for close races. The final outcome was 257 Democratic seats.

4.5. Covariation between states adds modest uncertainty

State polls are partially interdependent samples because they are conducted by a smaller group of polling organizations. This raises the likelihood that any systematic error will be shared by multiple states. One upper bound to the cumulative electoral effect of systematic error is the nominal 95% confidence band (the gray bands in Fig. 2). To test whether covariation was likely to contribute to the overall error, b was set to a range of $\Delta = [-1, +1]\%$ or $[-2, +2]\%$, and the resulting EV probability distributions were averaged over all values of Δ. This allows us to explore the question of whether polls are collectively biased by a constant amount, when the size and direction of the bias are unknown. The results for an August 2008 dataset are shown in Fig. 3.

All three cases showed the same median (298 EV) and mode (305 EV). With no covariation, the 68% confidence interval was [280, 312] EV, a width of 32 EV. With $\pm 1\%$ covariation, the confidence interval widened by 3 EV to [279, 314]. With $\pm 2\%$ covariation, the interval widened by 12 EV to [275, 319] EV. Thus, even when all state margins vary together perfectly, this results in only modest changes to the overall shape of the outcomes distribution.

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

Table 3
Performance comparisons in 2008 and 2012. Presidential predictions and results are listed for Barack Obama.

	FiveThirtyEight	Linzer (Votamatic)	InTrade	Polls alone (PEC)	Actual outcome
2008					
Presidential EV	348.5 EV	–	364 EV	353/364 EV	365 EV
Popular vote	52.3%	–	–	53.0%	52.9%
Senate	58–59 D	–	–	58–59 D	59 D
House	–	–	–	257 D	257 D
2012					
Presidential EV	313 EV	332 EV	303 EV	312 EV	332 EV
Brier score, Pres.win[a]	0.0083	0.0001	0.1170	0.0000	0.0000
Brier score, state win[a]	0.009	0.004	0.028	0.008	0.000
Senate close races	5/7	–	5/7	7/7	7/7
Brier score (30 races)[a]	0.045	–	0.049	0.012	0.000
Brier score, combined Presidential/Senate[a]	**0.023**	–	**0.037**	**0.009**	**0.000**

[a] Brier scores come from Table 5.2 of Muehlhauser and Branwen (2012), and are defined so that lower numbers indicate better performances. The 2012 Senate close races are listed in Section 4.8.

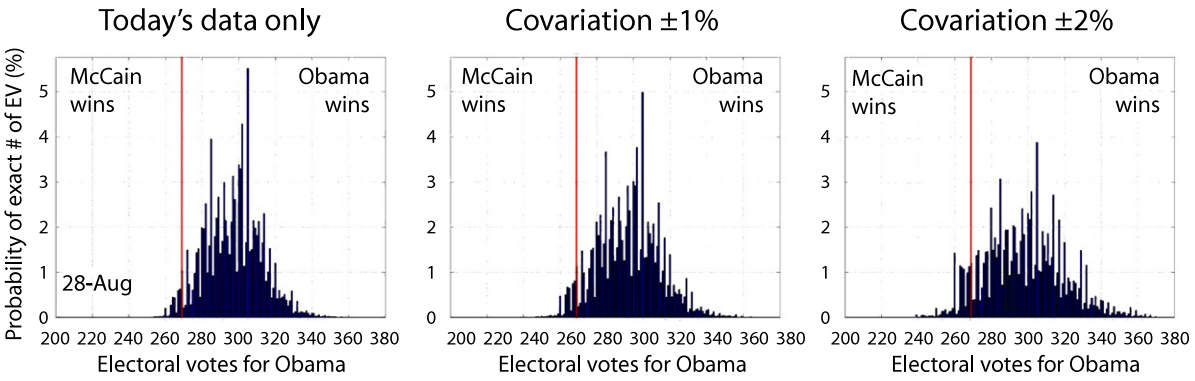

Fig. 3. Effects of covariation among state polls. The effect on (a) the uncorrected snapshot electoral vote estimator of adding a bias of (b) −1 to +1% or (c) −2 to +2% to state polls. The center of the distribution does not change, but its width increases modestly.

4.6. Obama vs. Romney 2012

Re-election races are generally thought of as a referendum on the incumbent. President Obama came into the general election campaign with a united Democratic party and a number of accomplishments, including the rescue of the auto industry and the passage of the Affordable Care Act. However, the economy was still weak and the opposition party was polarized and combative. Most fundamentals-based models gave the President a slight to moderate advantage for re-election (Graefe, Armstrong, Jones, & Cuzán, 2014; Montgomery et al., 2012).

Viewed as a whole from June 1 through to election day (Fig. 2(c)), the electoral history fluctuated around an equilibrium of Obama 312 ± 16 EV (mean ± SD), and a meta-margin of 3.0±1.2%. The distributions were not long-tailed (kurtosis = 2.7 for EV, 2.5 for the meta-margin, compared with 3 for a normal distribution). Thus, the race varied over about half the range of the 2004 election, and was notably stable.

The high time resolution of a state poll-based snapshot suggested that it might be possible to identify moments in time when opinion shifted suddenly (Fig. 4(a)–(c)). To quantify these turning points, I performed a breakpoint analysis via deviance minimization (O'Connor, Wittenberg, & Wang, 2005). For every date D from early August to the end of October, I calculated the sum-of-squares deviance over a 14-day interval, where the total deviance was calculated from averages within two subintervals: from $D-6$ to D, and from $D+1$ to $D+7$. The breakpoint score has a theoretical minimum value of zero, which can occur if the meta-margin is constant within each subinterval, but jumps up or down immediately after date D. This summed deviance was termed a breakpoint score (Fig. 4(d)). When the breakpoint score reaches a minimum, the meta-margin is most likely to have changed abruptly.

The breakpoint score reached a minimum value on five dates: August 14, September 2, September 23, October 4, and October 17. Each of these dates corresponded to a major campaign event: the addition of Rep. Paul Ryan (R) as Mitt Romney's running mate (the August 11th–17th news cycle, helping Romney), the Republican and Democratic National Conventions (August 27th–September 6th, helping Obama), the discovery of the 47% video (the September 17th–23rd news cycle, helping Obama), the first Presidential debate on October 3rd (helping Romney), and the second Presidential debate on October 16th (helping Obama). These tight temporal associations suggest that each campaign event triggered a discrete shift in the race (Fig. 4(d)). Thus, unlike a mixed polling/fundamentals-based approach, a polls-only approach is able to resolve notable campaign events to within one or a few days.

In particular, it should be noted that, because of its high temporal resolution, breakpoint identification does not require the meta-margin to be a perfectly accurate indicator of voter behavior. It has been suggested that

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

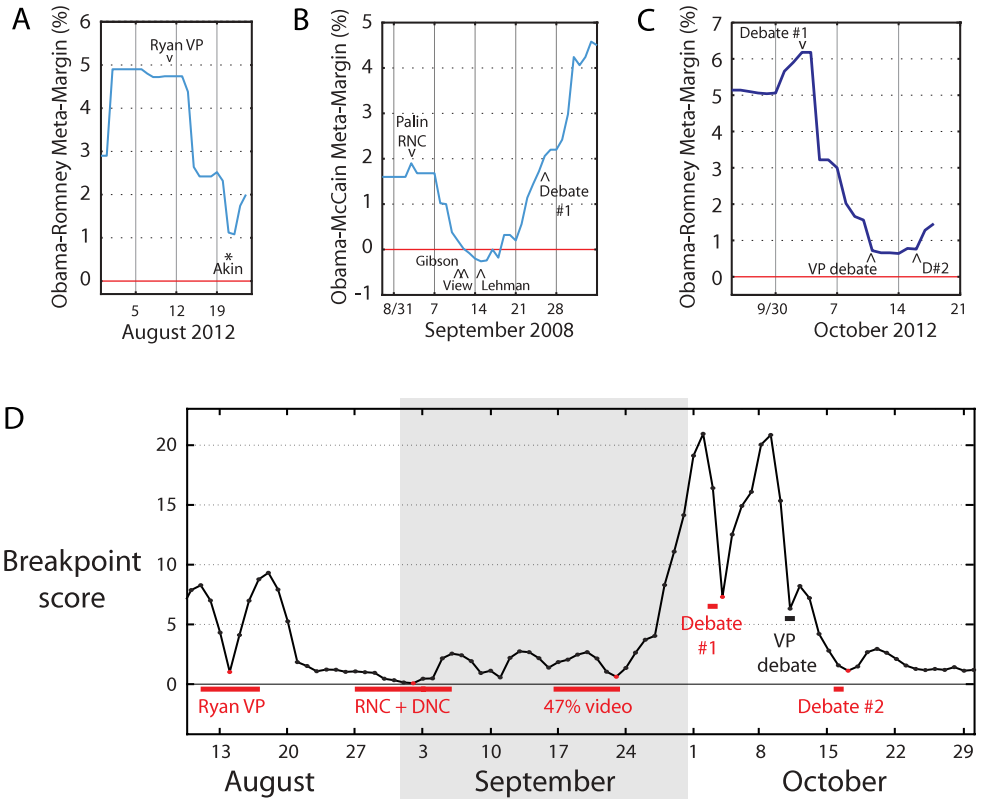

Fig. 4. Turning-point events in Presidential campaigns. An expanded view of significant campaign-moving events in 2008 and 2012, followed by subsequent events which are reported to have worked in the opposite direction. (a) Sarah Palin's (R) vice-presidential nomination acceptance speech at the Republican convention, followed by her interview with Charlie Gibson on ABC, John McCain's (R) appearance on The View, and the Lehman Brothers' collapse. (b) The announcement of the addition of Paul Ryan (R) as a vice-presidential nominee, followed by Rep. Todd Akin's (R) comment on "legitimate rape". (c) The first Obama–Romney presidential debate in 2012, followed by the Biden–Ryan vice-presidential debate and the second Presidential debate. (d) Breakpoints (red dots) indicate dates when a shift in opinion probably occurred. Breakpoints were defined as having the lowest breakpoint score (see the text) in a window extending seven days in both directions. Text labels indicate media events (including, where appropriate, a week-long news cycle) that were likely to be causal in driving the opinion shift. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

people answer polling questions differently later in the campaign (Enns & Richman, 2013; Gelman & King, 1993). However, these changes would be likely to be gradual. Any explanation for a shift in the meta-margin would have to account for the fact that, in many cases, breakpoints can be localized to a single day. For example, President Obama's performance in the first debate led to an immediate and massive shift in the way in which respondents answered polls. The parsimonious explanation is that, for a brief period, the debate pushed a substantial number of likely voters toward Mitt Romney.

4.7. A prediction with no fundamentals-based assumptions

Starting in 2012, PEC began to provide predictions. These were true predictions, but did not rely on economic and prior political conditions. Prediction was done using the same tool as was used to calculate the meta-margin and the effects of covariation. The prediction was constructed on the assumption that long-term movements in candidate preference moved uniformly in all states by an amount b, with b following a symmetric distribution with $\mu = MM$ and $\sigma = 2.2\%$. The parameter σ was estimated based on the movements of the meta-margin in the 2004 and 2008 races. Since the actual σ was 1.2% in 2012, this parameter was set conservatively.

The November prediction was plotted in the style of a hurricane strike zone, with the one-sigma band based on the parameter b (68% confidence interval) plotted in red, and a 95% confidence interval that included both long-term movement and pollster variations plotted in yellow (Fig. 2(c)). This random-drift prediction approach gave an Obama win probability of 90% in July.

To determine how quickly the shift b developed, I calculated the average change in the meta-margin for varying amounts of time from all dates in the 2008 general campaign season (Fig. 5). This quantity increased with a half-rise time of 20 days. Its time course was similar to a square root function, consistent with a random walk. Therefore, for short-term predictions as the election drew near, I modeled the movement in 2012 using $\sigma = 2.2 * \sqrt{(D/20)}$, where D was the number of days to the election. Under these assumptions, the Obama win probability increased to a maximum of 99.2% on election eve.

National polls could also be added as a Bayesian prior in order to inform an estimate of the national popular

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

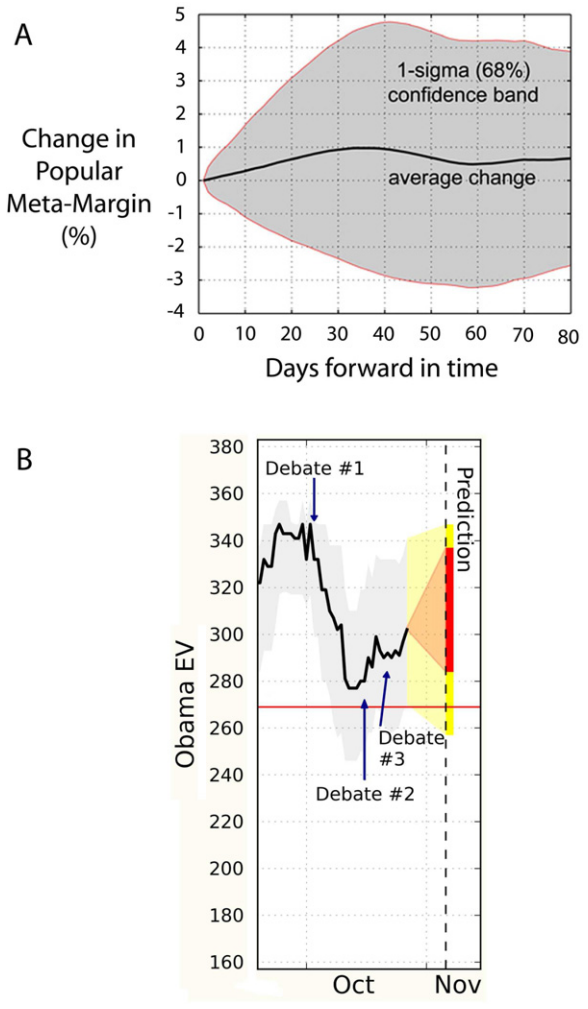

Fig. 5. A random-drift Bayesian prediction model for Presidential campaigns. (a) Average change in the meta-margin over the 2012 campaign season. (b) Application of the drift in (a) for making predictions. The red zone indicates the one-sigma range, while the yellow indicates the union of the two-sigma range and the 95% nominal confidence interval.

vote (Fig. 6). On the day before the election, the national poll median (Obama +0.0%) was assumed to predict the meta-margin as a t-distribution with $\sigma = 2.5\%$, a weak prior because of the substantial potential for systematic error. When combined with a state-polls-based prediction of Obama $+2.9 \pm 1.5\%$, the predicted popular-vote margin was Obama +2.4%, with a win probability >99.9%. The final two-party popular-vote margin was Obama +4.0%. Thus, state polls by themselves outperformed national polls in predicting the national popular vote.

4.8. Presidential coattails in the 2012 Senate race

Senate polls were analyzed using the same probabilistic algorithm as the EV tracker. The movement in this index was driven largely by seven close races: Missouri (D-i-McCaskill vs. R-Akin), Indiana (D-Donnelly vs. R-Mourdock), Massachusetts (D-Warren vs. R-i-Brown),

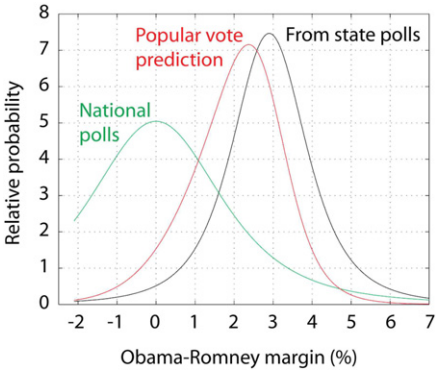

Fig. 6. Using state and national polls to predict the popular vote. National polls and the state-poll-based meta-analysis are combined to make a prediction of the national popular vote. The state-polls-only estimate performed better than the combined estimate.

Montana (D-i-Tester vs. R-Rehberg), North Dakota (D-Heitkamp vs. R-Berg), Virginia (D-Kaine vs. R-Allen), and Wisconsin (D-Baldwin vs. R-Thompson). PEC polling medians called the winner correctly in all seven races (Table 3).

Over time, the Senate seat-number tracking index (Fig. 7) moved up and down in parallel with the Presidential race. From mid-September to election day, the probability of a retained Democratic control stayed in the 80–99% range. A sharp dip in the Democratic/Independent seat count occurred in mid-August after the Ryan vice-presidential nomination, a steady and large increase occurred starting at the time of the Democratic convention, and a small decrease occurred after the first Presidential debate. Similarly to the Presidential EV tracker, the Republican convention led to little change in the Senate seat count, with, if anything, a slight movement toward Democrats.

These results indicate that Presidential and Senate preferences moved in tandem with one another, which is consistent with a coattail effect, i.e., similar party preferences at different levels of the ticket. However, the first Presidential debate had a relatively weaker effect on the Senate races than on the Presidential race, which suggests that the two levels are not always coupled equally.

5. Discussion

The principal conclusion of this study is that state polls by themselves, under the assumption that pollsters are accurate in the aggregate, are fully sufficient to make high-quality snapshots and predictions of the Presidential race. As early as Memorial Day, tracking and prediction can be done without the need for either corrections of individual pollsters or economic/political assumptions. Using statistical analysis alone, the meta-analysis combines polls to give a single snapshot with a temporal resolution approaching one day, and an accuracy equivalent to less than half a percentage point of difference in national support between the two candidates. Taken together, these qualities make the meta-analysis a sensitive indicator of the ups and downs of a national campaign — in short, a precise electoral thermometer.

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

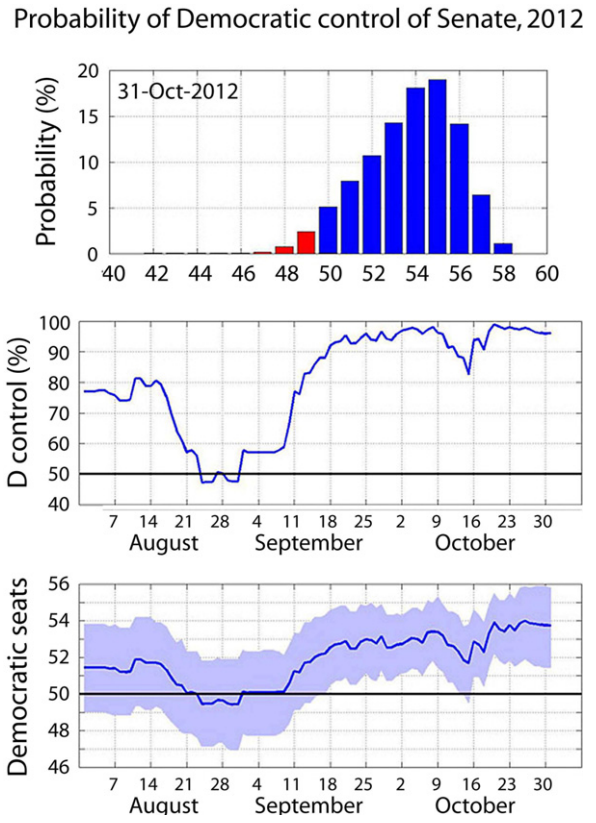

Fig. 7. Coattail effects in the U.S. Senate elections, 2012. Polling snapshot of Senate outcomes as a function of time, based the most recent available Senate data.

A post-election analysis (Muehlhauser & Branwen, 2012) has reviewed PEC's polls-only performance and found it to be significantly superior to other aggregators and the betting site InTrade, and nearly as good as the more complex Bayesian model from Votamatic (Table 3). This is made possible by the fact that pollsters show a wisdom of crowds effect in which their net bias is nearly zero. Enough state polls are available to enable a tracking of presidential races since 2000, when Ryan Lizza at *The New Republic* compiled state polls. On the day before the election, that compilation indicated that the outcome would hinge on Florida, as was ultimately the case. In 2004–2012, the state poll meta-margin came within an average of 1.6% of the national popular vote, with no sign errors (Table 1). The national margins in 2000–2012 have done worse, getting the sign of the popular-vote margin correct in only two years (2004 and 2008), and deviating from the actual outcomes by an average of 2.1%.

House-effect corrections of individual pollsters, as are done by aggregators such as FiveThirtyEight, appear to be unnecessary for the production of accurate predictions. To date, such corrections have not yielded much benefit in electoral-vote estimation (Table 3), though they are useful for statistical error analysis. In 2004, 2008, and 2012, the nominal confidence intervals of the EV estimator were wider than the event-related swings in each race. An accurate estimation of confidence intervals would require the removal of the contributions of house effects in individual polls before they are entered into the EV estimator.

The results here demonstrate that a model that is composed of uncorrected polls and random drift over time is fully sufficient for making highly accurate predictions. Therefore, if the goal is to predict the Presidential race or Senate seat counts during the election year, further assumptions appear not to add accuracy, and are therefore undesirable on grounds of parsimony. Logically, this suggests that, by the start of the campaign season, the information contained in those additional assumptions is already contained in state polls.

However, the additional assumptions still have useful applications that are not addressed by the meta-analysis. The meta-analysis does not address problems of missing data. In cases where polls are extremely sparse or unavailable, information about demographics or past voting patterns can be useful for interpolating results for specific races. As an example, FiveThirtyEight and Votamatic made accurate predictions in unpolled states in the Presidential race; but FiveThirtyEight incorrectly predicted Republican wins in the Montana and North Dakota Senate races, where poll medians correctly showed a Democratic lead.

The converse question arises: when do fundamentals contribute usefully to true prediction? It has been demonstrated (Abramowitz, 2008; Linzer, 2013) that economic and political variables have predictive value before a general election campaign, when polls are scarce. Once the season begins, opinion polls provide a direct measurement of opinion, at which point the problem becomes one of estimating how opinion will evolve over time. A true prediction properly done should not change much over time, as was seen in the work of Linzer (2013). A snapshot tracking the current state of the race does the converse. Adding random drift to the snapshot lacks an explanatory component, but has the advantage of generating a reliable forecast.

One way to incorporate fundamentals-based modeling while retaining the news power of the snapshot is to estimate the direction of the drift, going forward in time from the snapshot. For example, it should be possible to quantify how 2nd-quarter unemployment and the July-to-November poll *movement* are related, and with what distribution. In this manner, polling data at any moment in time could be used as a starting point for future projections.

Although national polls are inferior for presidential race prediction, they have the advantage of a high time resolution, due to their frequency. In contrast, the state-poll snapshot takes at least a week to equilibrate after a major campaign event. In the future, it may be possible to use national polling data to estimate day-to-day shifts in opinion (Fig. 8), and apply this to the EV estimator as a correction using the bias variable b, thereby achieving both accuracy and temporal sensitivity.

6. Conclusion

What is the future of poll aggregation? In addition to its news value, poll aggregation also has other applications. One is election integrity. In cases where substantial pre-election polling is available, fraud is made more difficult by the presence of concrete opinion data. A second application

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003

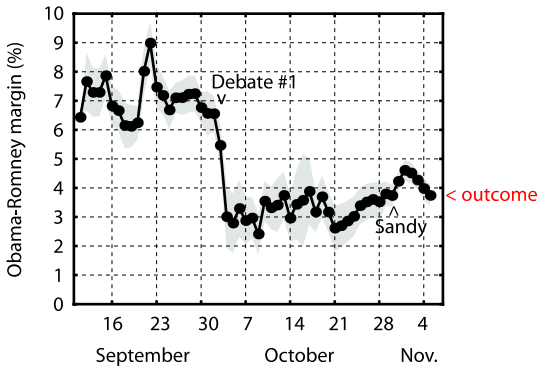

Fig. 8. Increased time resolution from the day-by-day averaging of national polls. National polling margins. Each available poll at Huffington Post/Pollster.com was distributed over the dates on which it was conducted and the average calculated. The time series was shifted so that the last day matched the actual popular vote outcome on election day. "Sandy" indicates Hurricane Sandy.

is resource allocation (Strömberg, 2002), both by candidate campaigns and by activist organizations. A third potential application is a reduction in media chatter concerning individual polls.

An open question for the future is whether poll aggregation will continue to perform well in the future. The answer depends in large part on the availability of accurate polling data. Economic tension exists between polling organizations, which release individual data points as a means of calling attention to themselves; news organizations, for which it is cheaper to run a poll than to pay a reporter for generating a story; and aggregators, who obtain far more accurate results by collecting many polls. Although one possible outcome is that fewer polls will be available, the effect on the meta-analysis would be minimal even if they were halved in number. Conversely, journalism might benefit from the weeding-out of low-information news stories about single polls. Ideally, this would clear the way for a more substantive coverage of political races.

Acknowledgments

I thank my collaborators Andrew Ferguson and Mark Tengi for establishing and maintaining the automated calculations and online presence of the Princeton Election Consortium, and Mark Blumenthal and colleagues at the Huffington Post/Pollster.com organizations for generously providing data feeds from 2008 to 2012. The methods described here have benefited from input, and in some cases code, from Alan Cobo-Lewis, Lee Newberg, Drew Thaler, and many others, including Rick Lumpkin, who performed the analysis for Fig. 7. Finally I thank my family, the Princeton Department of Molecular Biology, and the Princeton Neuroscience Institute for their support.

References

Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science and Politics*, *41*, 691–695.

Altman, D. (2014). Why is Nate Silver so afraid of Sam Wang? The Daily Beast, online, http://www.thedailybeast.com/articles/2014/10/06/why-is-nate-silver-so-afraid-of-sam-wang.html, October 6, 2014.

Banzhaf, J. F. (1965). Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review*, *19*, 317–343.

Enns, P. K., & Richman, B. (2013). Presidential campaigns and the fundamentals reconsidered. *Journal of Politics*, *75*, 803–820.

Erikson, R. S., & Wlezien, C. (2012). *The timeline of Presidential elections: how campaigns do (and do not) matter*. Chicago: University of Chicago Press.

Forelle, C. (2004a). For math whizzes, the election means a quadrillion options. *Wall Street Journal*, October 26, A1.

Forelle, C. (2004b). Winner at picking electoral vote. *Wall Street Journal*, November 4, D9.

Gelman, A., & 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.

Gelman, A., Silver, N., & Edlin, A. (2010). What is the probability your vote will make a difference? *Economic Inquiry*, *50*, 321–326.

Graefe, A., Armstrong, J. S., Jones, R. J. J., & Cuzán, A. G. (2014). Accuracy of combined forecasts for the 2012 presidential elections: the PollyVote. *PS: Political Science and Politics*, *47*, 427–431.

Jackman, S., & Blumenthal, M. (2013). Using model-based poll averaging to evaluate the 2012 polls and pollsters. In *AAPOR 68th annual conference*.

Jones, R. E., Jr. (2008). The state of presidential election forecasting: the 2004 experience. *International Journal of Forecasting*, *24*, 310–321.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2013). Motivated numeracy and enlightened self-government. http://dx.doi.org/10.2139/ssrn.2319992.

Lewis-Beck, M. S., & Tien, C. (2008). Forecasting presidential elections: when to change the model. *International Journal of Forecasting*, *24*, 227–236.

Linzer, D. A. (2013). Dynamic Bayesian forecasting of Presidential elections in the states. *Journal of the American Statistical Association*, *108*, 124–134.

Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Ensemble predictions for the 2012 US Presidential election. *PS: Political Science and Politics*, *45*, 651–654.

Muehlhauser, L., & Branwen, G. (2012). Was Nate Silver the most accurate 2012 election pundit? *Center for Applied Rationality*, http://rationality.org/2012/11/09/was-nate-silver-the-most-accurate-2012-election-pundit/.

Noonan, P. (2012). Monday morning. *Wall Street Journal*, online, http://blogs.wsj.com/peggynoonan/2012/11/05/monday-morning/, November 5, 2012.

O'Connor, D. H., Wittenberg, G. M., & Wang, S. S.-H. (2005). Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 9679–9684.

Poor, J. (2012). George Will predicts 321-217 Romney landslide. *Daily Caller*, http://dailycaller.com/2012/11/04/george-will-predicts-321-217-romney-landslide/.

Soumbatiants, S. R. (2003). *Forecasting the probability of winning the U.S. Presidential election*. Doctoral thesis, University of South Carolina.

Soumbatiants, S., Chappell, H., & Johnson, E. (2006). Using state polls to forecast U.S. Presidential election outcomes. *Public Choice*, *123*, 207–223.

Strömberg, D. (2002). *Optimal campaigning in Presidential elections: the probability of being Florida*. Seminar Paper No. 706, Institute for International Economic Studies, Stockholm University.

Tracy, M. (2012). Nate Silver is a one-man traffic machine for The Times. *The New Republic*, November 6, 2012.

Please cite this article in press as: Wang, S. S.-H., Origins of Presidential poll aggregation: A perspective from 2004 to 2012. International Journal of Forecasting (2015), http://dx.doi.org/10.1016/j.ijforecast.2015.01.003