2008-04-23

The Optimal Weighting of Pre-Election Polling Data

Gregory K. Johnson
Brigham Young University - Provo

THE OPTIMAL WEIGHTING OF PRE-ELECTION POLLING DATA

by

Gregory K. Johnson

A selected project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

August 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a project submitted by

Gregory K. Johnson

This selected project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	William F. Christensen, Chair
Date	Scott D. Grimshaw
Date	C. Shane Reese

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the selected project of Gregory K. Johnson in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date	William F. Christensen Chair, Graduate Committee

Accepted for the Department

Date	Scott D. Grimshaw Graduate Coordinator

Accepted for the College

Date	Thomas W. Sederberg Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

THE OPTIMAL WEIGHTING OF PRE-ELECTION POLLING DATA

Gregory K. Johnson

Department of Statistics

Master of Science

Pre-election polls are used to test the political landscape and predict election results. The relative weights for the state-level data from the 2006 U.S. senatorial races are considered based on the date on which the polls were conducted. Long- and short-memory weight functions are developed to specify the relative value of historical polling data. An optimal weight function is estimated by minimizing the discrepancy function between estimates from weighted polls and the election outcomes.

ACKNOWLEDGEMENTS

I wish to expressly thank Dr. William Christensen for his efforts, and ongoing interest in my behalf. The same is true for the faculty in the BYU Statistics Department. As a group, they have provided me truly exceptional support, assistance, and consideration in obtaining my master's degree. I remain indebted to them. I would also like to thank my dear wife Karen who suffers me so gladly. I love her.

CONTENTS

FIGURES

Figure

1. INTRODUCTION

For decades, politicians and the like have used polling data to predict election results. As campaign budgets have soared, so has the frequency and sophistication of conducting pre-election polls. Subsequently, accurately analyzing and interpreting the results of these polls has become increasingly more important. The natural consequence of increased frequency and sophistication is an increased difficulty in the analysis and proper interpretation of results of polling activities. As an illustration of the difficulty of proper interpretation of results of polls, Carl Bialik reported in the Wall Street Journal (2008) that some pollsters have merely averaged the results of polling data. The aim of this averaging is to offset conflicting results, to control for competing interests, and to achieve a more accurate synopsis of the political landscape. Mr. Bialik observes

> Among the pitfalls: Polls have different sample sizes, yet in the composite, those with more respondents are weighted the same. They are fielded at different times, some before respondents have absorbed the results from other states' primaries. They cover different populations, especially during primaries when turnout is traditionally lower. It's expensive to reach the target number of likely voters, so some pollsters apply looser screens. Also, pollsters apply different weights to adjust for voters they've missed. And wording of questions can differ, which makes it especially tricky to count undecided voters. Even identifying these differences isn't easy, as some of the included polls aren't adequately footnoted.

The statistical issues associated with simple averaging of polls are clearly laid out in Mr. Bialik's comments. To be precise, the statistical problems associated with simple averaging of polls is the lack of accounting for

1. different sample sizes among polls,
2. different populations of interest (as evidenced by poll time, and sampling frame),
3. different sampling weights provided by polling organization, and
4. different question wording which induces a bias.

Poll result synthesis based on simple averaging is therefore an inadequate approach. This is not to say that there is little value in polling data, rather the opposite: all polling results are of value, it is simply a question of the degree to which polls are considered valuable. A more appropriate analysis would seek to weight different polls, accounting for some of the effects mentioned above.

The purpose of this study is to determine optimal weight functions for polling results which minimize the discrepancy between polls and the actual election results. To better address these weights, we explore different options for weights using state-level pre-election polling data from the 2006 senatorial races and their subsequent results (31 states qualified for our study). The relative value of the polls was based on the dates on which they were conducted. Two weight functions were considered, short memory and long memory, providing more weight to the more recent polls. The remainder of this project explores the construction of the optimal weight functions. More specifically, Chapter 2 contains a review of literature related to weighting function construction and poll synthesis. Chapter 3 presents the short and long memory weight functions developed

and the results of these two different weight functions on the data from the 2006 senatorial races. In Chapter 4 the relative merits of the two different approaches is discussed and conclusions are drawn.

2. LITERATURE REVIEW

For decades politicians, campaigns, political pundits, reporters, and academics have used polling data to predict election outcomes. As congressional campaigns become more sophisticated at measuring voter opinions and local media have more complete coverage of state races, there has been an increase in the frequency, sophistication, and publication of pre-election polls. Subsequently, accurately analyzing and interpreting the results of these polls has become increasingly more important and more difficult.

There are many concerns with embracing a single poll to predict election outcome. First, polls are, at best, snapshots of voter opinion. They are frequently reported weekly for national races or monthly for statewide contests and are often released on weekends to coincide with the Sunday political news cycle or leading into upcoming primary votes. Weekly sampling is based on the assumption that voter opinions change often, which is possible when significant events occur or political stumbles from the candidate or campaign change the message. However, most elections are stable with a constancy to message and strategy, and it seems reasonable to combine the information in a thoughtful way.

One approach is to create a regression model to predict the national popular vote. Polling data can be incorporated as an explanatory variable, although how to use past polls instead of simply the most recent poll requires more attention. Some of the explanatory variables describe the nature of the campaign and the amount of underlying partisan support. Examples include an indicator variable for incumbency, the current president's approval rating, number of party delegates, strength of third-party

challengers, and measurements of the national economy. One of the challenges is identifying explanatory variables that measure voter interest. Examples include historical voter turnout, degree of partisanship, satisfaction with education, defense, and other issues. To demonstrate this approach, consider Campbell (1992), where 16 explanatory variables were used to predict presidential outcome by September of an election year. These models can be modified to apply to state races. Even these models rely on polling data. Brown and Chappell (1999) found that poll data dominates the optimal forecast when compared to models with only explanatory variables based on historical election fundamentals.

Other models attempt to define the characteristics of likely voters. One of the flaws in polls is that people are usually surveyed by phone. While it is possible to ask if a person plans to vote, the answer is considered biased since most people in a survey believe it reflects poorly on their citizenship to admit to not voting. Some pollsters ask a series of screening questions to identify likely voters. Some are direct questions, such as "are you a registered voter" or "who did you vote for in the last congressional race." Some are indirect questions, where they ask questions you would need to know if you had voted such as "how long did you wait in line to vote last time" or "what time of day did you last vote." Another approach is to develop profiles of likely voters and weight the sample data to reflect the population. For example, pollsters will develop demographic groups or partisan groups and estimate the expected voter turnout.

However, these models do not address the issue of principal interest, namely who will actually win the presidency. While the popular vote and the electoral vote often agree, Al Gore takes little comfort for winning the popular vote in the 2000 U.S.

presidential election. Although national popular opinion during U.S. presidential races is most commonly measured and discussed in the media, the U.S. presidential election is based on the electoral college, in which each state has a number of electors equal to the number of its U.S. representatives. Additionally, the District of Columbia acts as a "state" with a number of electors proportional to its population, but not exceeding the number of electors assigned to any of the states. The people in each state vote for the state-level electors who then vote for a presidential candidate, with most states using a winner-take-all policy for casting votes in the electoral college. Thus, although much of the media attention during election years focuses on polls tracking popular support for the major candidates, the complicated role played by the electoral college in this multistage election process must be accounted for in order to address the issue of winning the presidency.

Bialik (2008) reports that some pollsters have merely averaged the results of polling data in an attempt to offset conflicting results and to achieve a more accurate synopsis of the political landscape. Current practice is described by Mark Blumenthal, a former Democratic pollster and co-founder of Pollster.com, as not optimal, but "lets hope that by combining them were getting some better version of the truth" (qtd. in Bialik 2008). This naive approach to combining polls from different days ignores different sample sizes. Sophisticated polling asks questions and applies sample weights that allow survey respondents' opinions to be portrayed as likely voters. Different pollsters use different filtering questions to identify partisan voters. Bialik notes that identifying and counting undecided voters is particularly challenging. Unfortunately, the details of a pollsters sample and operating procedure is not adequately disclosed. Without technical

descriptions it is difficult to provide a thoughtful method to combine information from different political polls.

Simply averaging polls is also a poor choice. This is not to say that there is little value in polling data, rather the opposite: all polling results are of value, it is simply a question of how much. Christensen and Florence (2008) describe a simulation-based approach (either frequentist or Bayesian) to answering election outcome questions that rely on combining polls.

Historically, one of the main challenges associated with forecasting election outcomes has been the lack of state-level-pre-election poll data (Cohen 1998), but opinion polls are now easily accessible on the Internet. For example, in 2004, state-level poll data for all 50 states and the District of Columbia were available from several web pages such as the LA Times website (where most of the data for these analyses were obtained). Although pre-election polling data are inevitably awed, they can still provide much insight about national and regional trends.

Political scientists who study elections have noted that presidential pre-election polling data may not be useful until at least early September after the two parties' national conventions. The analyses of the 2004 presidential election discussed use state-level opinion poll updates 12 different times beginning 12 October 2004 and ending 2 November 2004, the day before the election. During the 22-day window with poll results, some states had no new updates while others had as many as ten. With the beginning 12 October 2004 data, polls are assumed to be taken on that day even though some polls may have been older. Multiday polls were treated as if the data were gathered on the day the poll was reported.

3. METHODS AND ANALYSIS

In this chapter, we consider the optimization of poll weights for prediction of a single-stage election outcome. That is, we consider a data scenario similar to that observed in Christensen and Florence (2008). We are interested in predicting the actual percentage of a population voting for a specific candidate. Specifically, we consider each state separately and determine what weighting scheme within a class of weights will yield the best estimate of the actual percentage voting for a candidate.

For this exploration, we use the election poll data obtained prior to the November 2006 U.S. Senate races. Prior to this election, the Republican Party was in the majority in the U.S. Senate but several GOP senators were defending hotly contested seats. For most states, the polling data used in these analyses were gathered on 1 August 2006 or later. The exceptions were Massachusetts, Mississippi, and Wyoming, where the only polls available before the last week of the campaign were conducted prior to August 1. For each date between 1 August 2006 and the election on November 7, Christensen and Florence (2008) predicted the outcomes of each race and also predicted the likelihood of a change in majority party. For each prediction, the election polling data was formulated in one of three ways: (1) using only the latest poll, (2) combining all of the responses from all previous polls, and (3) weighting the responses from previous polls, with decreasing weights for older polls. They consider two different weighting functions—one that gives the estimator a long memory of the past polls and one that gives a short memory. The general form of the weight function is:

$$w(t;h,f) = \min\left(1 - \frac{t}{2h}, f\right), \tag{1}$$

where t is the number of days since the poll was carried out, h is the "half-life" of the function, and f is its "floor." To define the facets of this function, consider the "long memory weight function" defined by Christensen and Florence (2008) as follows:

$$w(t;35,0.2) = \begin{cases} 1 - \frac{t}{70}, & t \leq 56 \\ 0.2, & t > 56 \end{cases} . \quad (2)$$

This weight function is illustrated in Figure 3.1. Note that the weight function implies that a respondent to a poll that is 35 days old will have a weight equal to half that of a respondent in a poll released today. Thus, we reference the slope of this line with the function's "half-life" of 35 days. The interpretation of the weight function for $t \leq 56$ is that polling data has decreasing utility as it ages. The other parameter governing this class of weights is the minimum or "floor" weight. In the sample weight function given above, the floor is equal to 0.2. That is, at 56 days old, a poll's respondents will have a weight of 0.2, but will then decrease no more as it continues to age. The interpretation of the weight function for $t \geq 56$ is that the utility of polling data always retains some minimal level of value, regardless of age.

Christensen and Florence (2008) also use a "short-memory" weight function defined by

$$w(t;7,0.05) = \begin{cases} 1 - \frac{t}{14}, & t \leq 13 \\ 0.05, & t > 13 \end{cases} . \quad (3)$$

The half-life for the short-memory indicates that a poll has lost half of its utility by the time it is one week old. The floor value of 0.05 is also much smaller than in the long-memory weight function, indicating that in every respect, the estimator using this weight

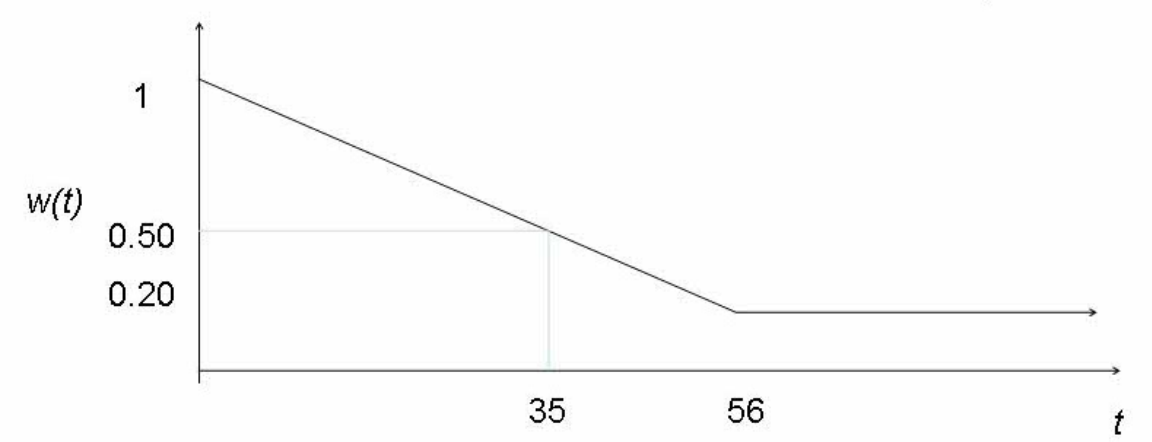

Figure 3.1. Illustration of weighting function used by Christensen and Florence (2008) when weighting polls.

will draw only minimally on older polling data. Figure 3.2 compares the nature of the long- and short-memory weight functions.

Note that for the predictions made in Christensen and Florence (2008) between August 1 and 6 November 2006, we cannot evaluate the accuracy of our state-by-state or overall predictions because there is no "ground truth" against which we can compare. However, the prediction on our final day can be compared to the actual results on 7 November 2006. That is, we cannot evaluate the optimality of our weights for August data when predicting voter behavior on September 1, but we can evaluate different weighting schemes for August-through-November data when predicting election results for November 7.

In this section, we consider the class of weights illustrated in equation (1) and identify the optimal weight function for each of the 31 senate races we were tracking. Additionally, we are interested in recommending one "all-purpose" weight function that can be used for future poll tracking of this nature. It may not seem optimal to use an all-purpose weight function when tracking a race for which we can obtain state-specific

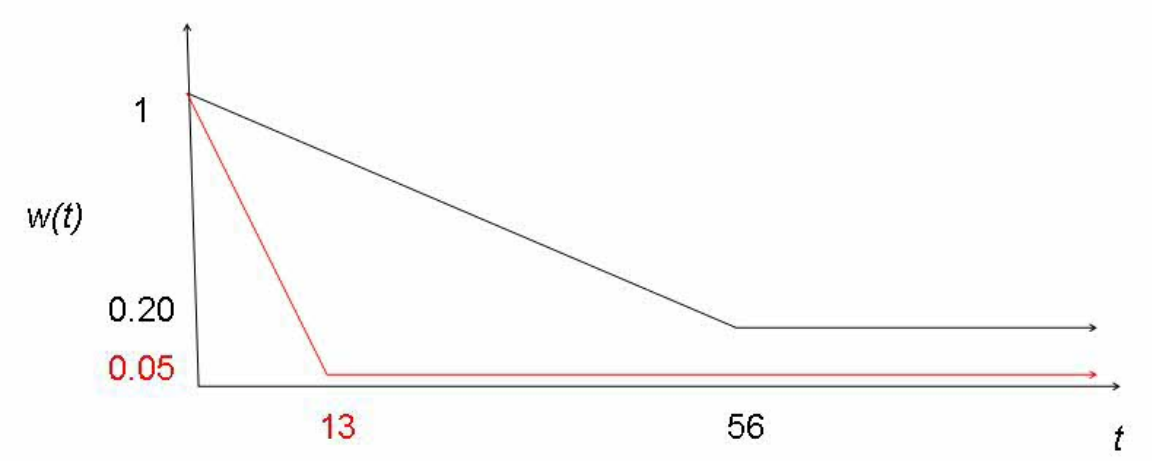

Figure 3.2. Long-memory weight function from equation (2) in black and short-memory weight function from equation (3) in red.

optimized weights. For example, we can obtain an optimized weight function for Tennessee based on the 2006 Senate data and then use that specific weight for the 2008 Senate race in Tennessee. However, it is also plausible that the 31 estimated optimal weight functions obtained from each of the Senate data sets in 2006 represents a distribution of estimates for some universal weight function. Under this assumption, we can simultaneously use the 31 data sets from 2006 to posit a function that is best for future use in an overall sense.

3.1 Optimal Weight Function for a Specific State

Consider the m_i pre-election polls for a given state, with poll ages $\{t_{i1}, \ldots, t_{im}\}$, poll sample sizes $\{n_{i1}, \ldots, n_{im}\}$, and Republican preference counts $\{r_{i1}, \ldots, r_{im}\}$. Because we do not want our calculations influenced by potential voters who are undecided or voting for third-party candidates, our "sample size" for these calculations is actually the sum of the Democratic preference count and the Republican preference count (i.e., $n_{ij} = d_{ij} + r_{ij}$). We consider all possible combinations of the half-life (h) in $H=\{1, 2, \ldots, 50\}$

with floor (f) in F={0.0001, 0.01, 0.02, …, 0.50}. For each pair (h, f), we calculate the estimate of the proportion voting Republican in state i (among all persons voting Republican or Democrat) with

$$\hat{\pi}_i(h, f) = \frac{\sum_{j=1}^{m} w(t_{ij}; h, f) \times r_{ij}}{\sum_{j=1}^{m} w(t_{ij}; h, f) \times n_{ij}} \quad . \tag{4}$$

We consider the optimal weight function for the state to be $w(t; h_i^o, f_i^o)$, where

$$(h_i^o, f_i^o) = \arg\min_{h \in H, f \in F} | \hat{\pi}_i(h, f) - \pi_i | \tag{5}$$

and π_i is the actual proportion voting for the Republican candidate in state i (among all persons voting for either the Republican or the Democrat).

3.2 Overall Optimal Weight Function

Our task is then to choose an "overall" weight function that in some sense best predicts the vector of Republican preference proportions for all 31 states ($\boldsymbol{\pi} = \pi_1, \ldots, \pi_{31}$). A simple rule for choosing the optimal values of h^o and f^o in the overall weight function $w(t; h^o, f^o)$ is to minimize the discrepancy function

$$D_1(h, f) = \frac{1}{31} \sum_{i=1}^{31} | \hat{\pi}_i(h, f) - \pi_i | \tag{6}$$

so that

$$(h^o, f^o) = \arg\min_{h \in H, f \in F} D_1(h, f) \quad . \tag{7}$$

The problem with the rule in (7) is that it penalizes estimation errors equally across states. So, if we estimate a state to yield 70% Republican vote instead of an actual

value of 75%, this has equal impact on the discrepancy measure as if we estimate a state to yield 47% Republican vote instead of an actual value of 52%. In order to give greater weight to the close races, we could weight each term in the sum found in (6) using some measure of tightness. In this study, we use the number of published polls for a state (m_i) as a measure of a race's tightness to obtain the discrepancy function

$$D_2(h, f) = \frac{1}{31}\sum_{i=1}^{31}\left(| \hat{\pi}_i(h, f) - \pi_i | \times m_i \right) . \quad (8)$$

Then, our optimal function is defined using

$$(h^o, f^o) = \arg\min_{h \in H,\, f \in F} D_2(h, f) . \quad (9)$$

Thus, states generating the most election coverage by pollsters (e.g., battleground states) will have the largest influence in selecting the optimal weight function. Alternatively, one could weight by the closeness of $\hat{\pi}_i$ to 0.50 in the discrepancy function, as in

$$D_3(h, f) = \frac{1}{31}\sum_{i=1}^{31}\left(\frac{| \hat{\pi}_i(h, f) - \pi_i |}{| \hat{\pi}_i(h, f) - 0.50 |} \right)$$

or

$$D_4(h, f) = \frac{1}{31}\sum_{i=1}^{31}\left(| \hat{\pi}_i(h, f) - \pi_i | \times \left[0.5 - | \hat{\pi}_i(h, f) - 0.50 | \right] \right) .$$

Figure 3.3 gives a plot showing the values of $D_2(h, f)$ for all possible combinations of the half-life (h) in H={1, 2, …, 50} with floor (f) in F={0.0001, 0.01, 0.02, …, 0.50}. Note that the overall optimal weight function is $w(t; h^o, f^o) = w(t; 20, 0.0001)$. That is, the weight function that minimizes $D_2(h, f)$ in equation (8) is one that gives polls a half-life of 20 days and a floor value of essentially zero. (We do not set the value of the floor at zero because there are some states for which the most recent poll

may be older than $2h$.) The optimal weight function for each of the ten closest senate races is also denoted on the plot. Note that 6 of the 10 closest states (and 15 out of 31 states in total) use virtually no weight for polls older than $2h$ (i.e., $f^o = 0.0001$). We recommend the weight function $w(t; h^o, f^o) = w(t; 20, 0.0001)$ for general use in future work predicting election outcomes from pre-election polls.

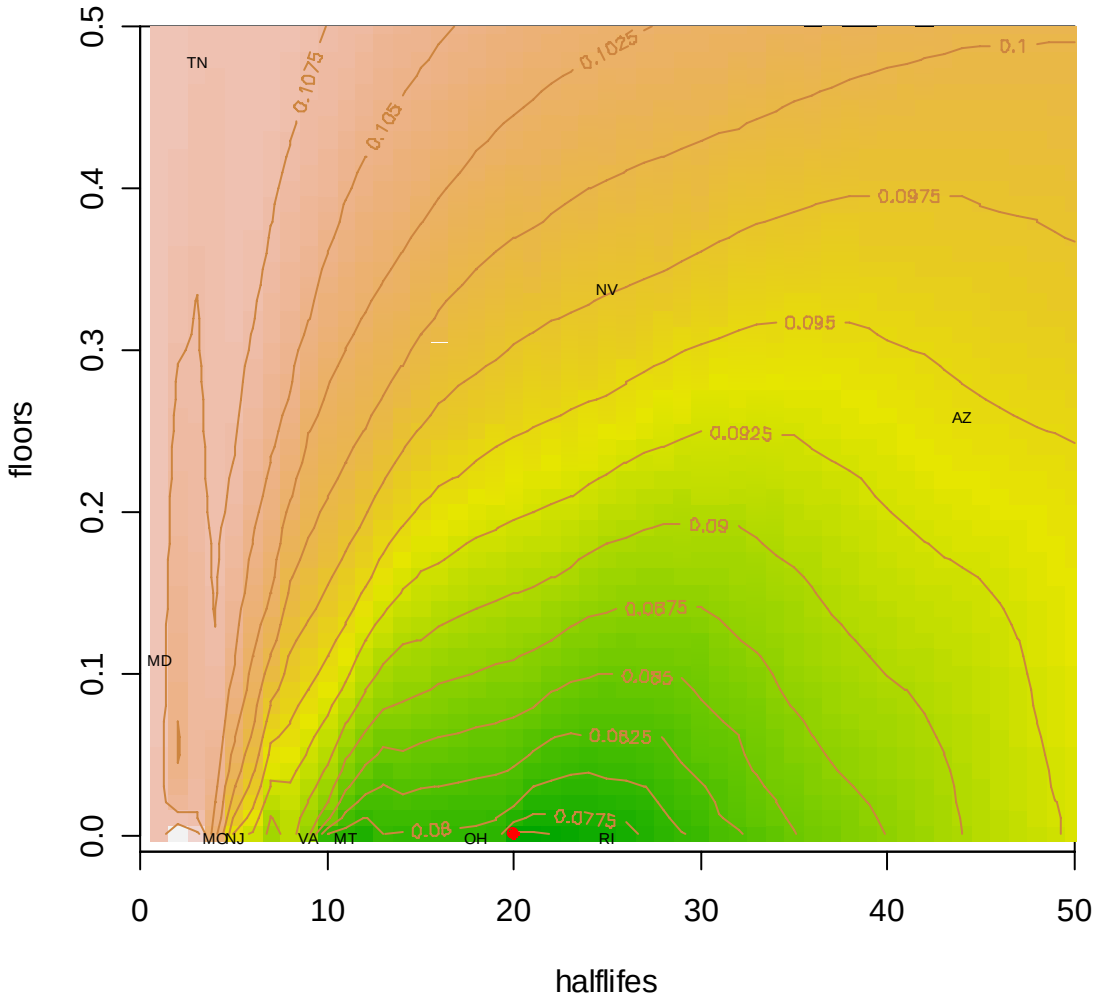

Figure 3.3. Values of $D_2(h, f)$ for combinations of half-life (h) in $H=\{1, 2, \ldots, 50\}$ with floor (f) in $F=\{0.0001, 0.01, 0.02, \ldots, 0.50\}$. The red dot indicates the minimum value of $D_2(h, f)$ with $(h^o, f^o) = (20,\ 0.0001)$. Optimal values for the weight functions associated with the ten closest senate races are denoted with state abbreviations.

4. CONCLUSION

In summary, the purpose of this study is to provide better means by which polling data may be utilized. Polls are increasingly more expensive and relied upon. Simply averaging the polls does not account for differences in sample sizing, populations of interest, pollsters, question wording, and so forth, and therefore can skew interpretations.

A total of 2,550 weight functions are considered, each having a piecewise linear form. The overall optimal weight function for these data is determined based on the notion that the specific function for each state is a random realization from an overall distribution with common "average" shape. With this assumption, it is determined that a poll has a half-life of 20 days, and a floor value of essentially zero, meaning that a poll loses its value within 10 days and has no value thereafter.

Our approach for choosing the optimal weight function gives a much larger influence to the states with the closest races. If one is interested in giving an equal influence to all races, a different optimal weight function would be determined.

BIBLIOGRAPHY

Bialik, C.N., "Election Handicappers are Using Risky Tool: Mixed Poll Averages," The Numbers Guy, *Wall Street Journal*, sec. B1, February 15, 2008.

Brown, L.B., and Chappell, H.W. Jr. (1999), "Forecasting Presidential Elections using History and Polls," *International Journal of Forecasting*, 15, 127–135.

Campbell, J.E. (1992), "Forecasting the Presidential Vote in the States," *American Journal of Political Science*, 36, 386–407.

Christensen, W.F., and Florence, L.W. (2008), "Predicting Presidential and Other Multistage Election Outcomes Using State-Level Pre-Election Polls," *The American Statistician*, 62, 1–10.

Cohen, J.E. (1998), "State-Level Public Opinion Polls as Predictors of Presidential Election Results: The 1996 Race," *American Politics Quarterly*, 26, 139–159.