Gov 2002 - Section 9 - Polling and Political Research in an Uncertain Era: A Case Study

Connor Jerzak

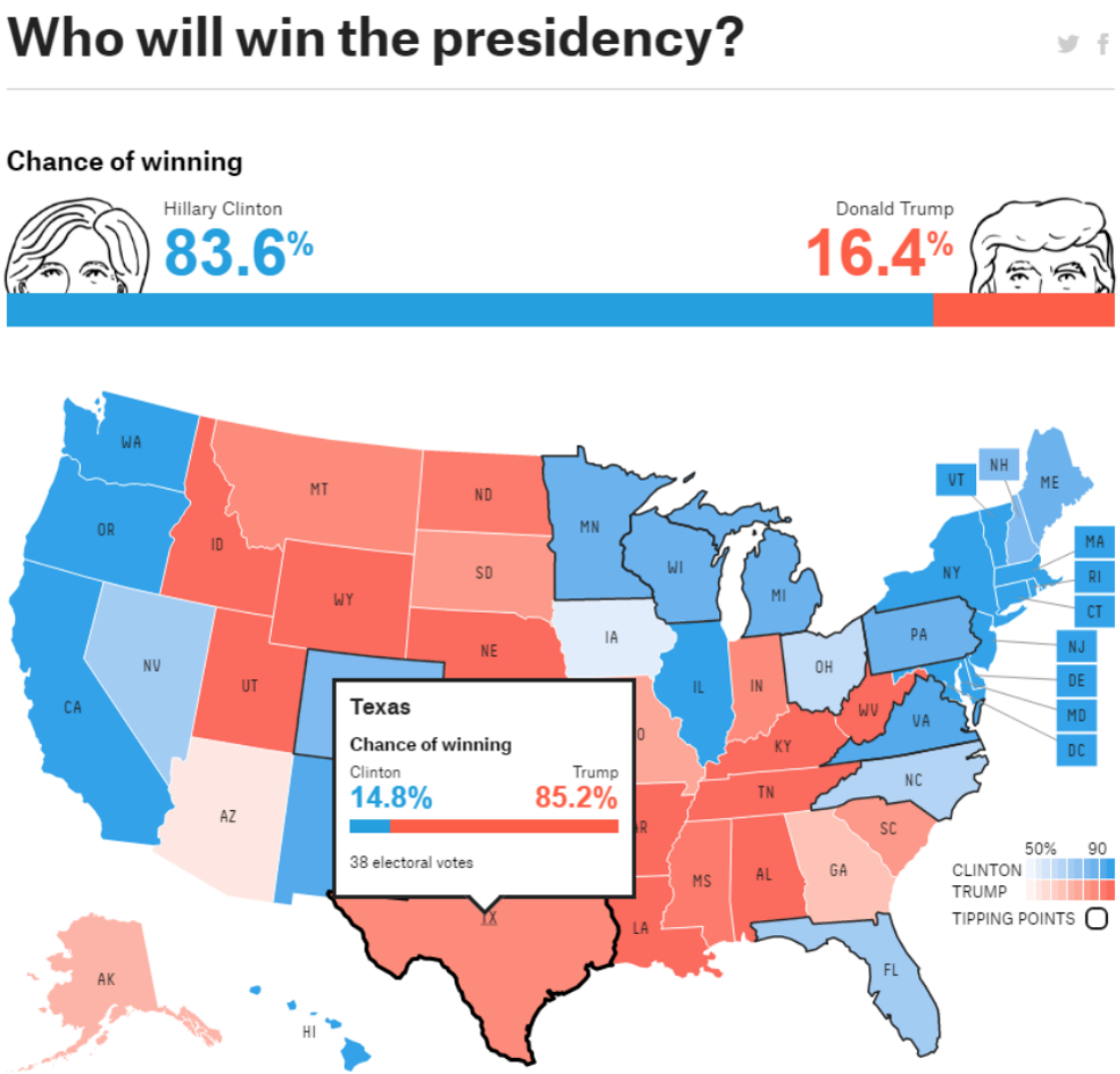
Harvard University, Department of Government

November 9, 2016

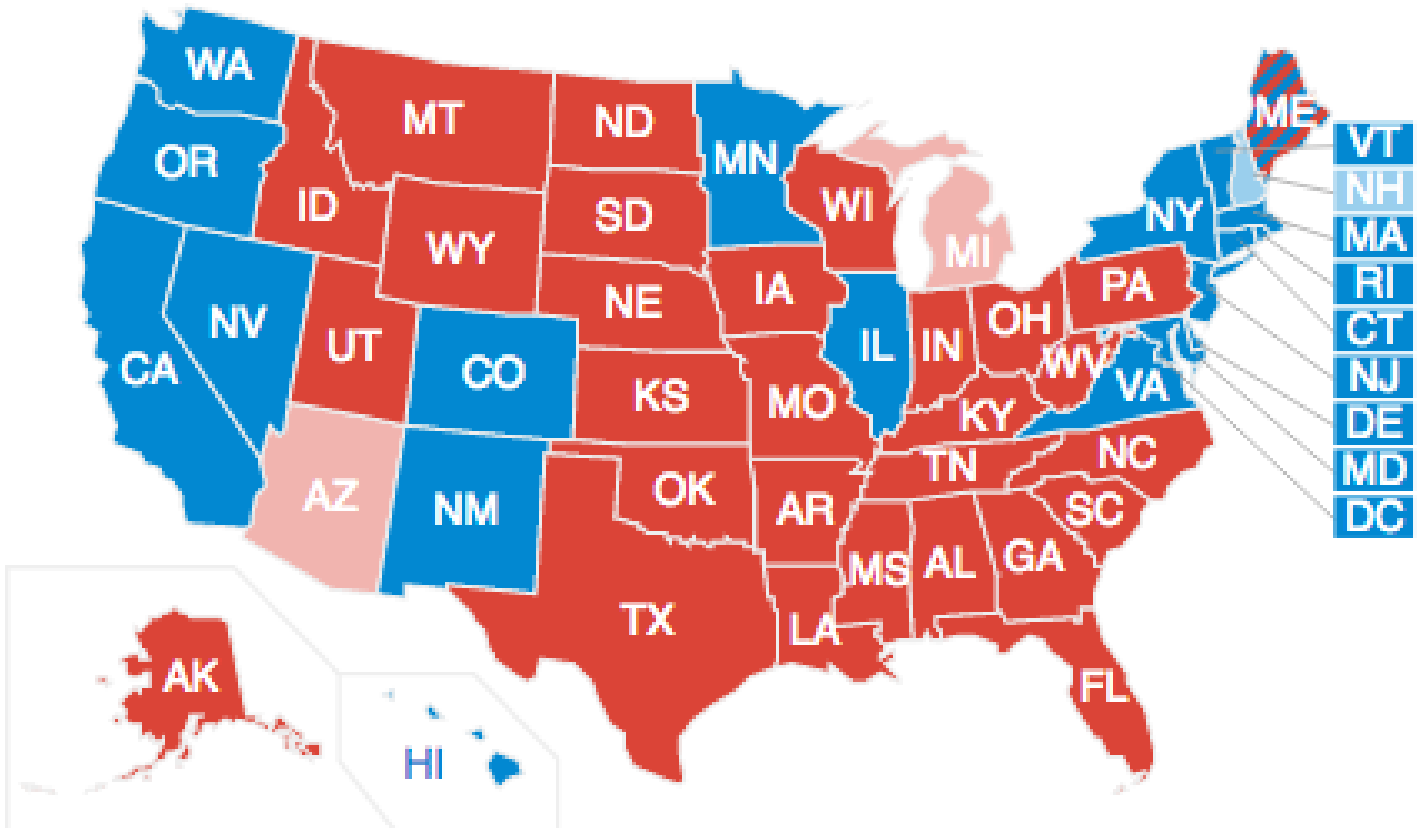
Introduction

- Quantitative models predicted a victory by Hillary Clinton over Donald Trump in the 2016 presidential election with a high degree of certainty:
  1. FiveThirtyEight: $\approx 75\%$
  2. Huffington Post: $\approx 95\%$
  3. New York Times: $\approx 85\%$

Predicted


Who will win the presidency?
Chance of winning
Hillary Clinton
83.6%
Donald Trump
16.4%
Texas
Chance of winning
Clinton
14.8%
Trump
85.2%
38 electoral votes
50%
90
CLINTON
TRUMP
TIPPING POINTS
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
VT
NH
ME
MA
RI
CT
NJ
DE
MD
DC
AK
HI

Actual


WA
OR
CA
NV
ID
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
ME
VT
NH
MA
RI
CT
NJ
DE
MD
DC
AK
HI

This kind of occurrence has become increasingly common.

1. Major polls showed Clinton with a large lead over Bernie Sanders in the Michigan Democratic primary of 2016.
2. Poll averages missed the final outcome of the 2016 Brexit vote by 4%.
3. Polling for the $2006, 2010$ midterm elections, as well as the 2012 presidential election, understated support for Democratic candidates. The situation was the opposite for the elections of $2012, 2014, 2016$.
4. Polling understated support for the Likud Party in the 2015 Israeli elections.
5. ... etc.

Today

- What can political scientists learn from this apparent breakdown of quantification?
  1. Insights for measurement.
  2. Insights for analysis.
- All quantitative research boils down to just this - **measurement** and **analysis**. We must understand the breakdown of polling predictions — and what it means for our own research — in these terms.

Insights for measurement

- Surveys are the input data for many election prediction systems and for much political science research. They are highly imperfect.
  1. **Non-response bias**.
  2. **Social desirability bias & cheap talk**.

Non-response bias

- Response rate to polling agencies can be as low as 2-9% according to Pew Research (down from 36% in 1997). People who the agencies reach may be systematically different from those who they do not reach.
  1. Better-educated, more politically interested people are more likely to take surveys dealing with politics.
  2. Non-response rates are growing across the advanced democracies. Consequence of caller cell phone, ID technologies, as well as concerns about privacy.
  3. Related to idea of *differential non-response*. Swings in polls can reflect different levels of excitement about a candidate/issue (and thus who is willing to answer a political survey), not genuine changes in attitudes.
  4. Weighting can reduce but not eliminate this problem.
     - Why not?

Weighting cannot always cure the non-response disease

Assume I have three apples and three oranges. I want to know the true overall weight.

$$O_1, A_1, O_2, A_2, O_3, A_3.$$

- I observe: O_1, A_1, O_2, A_2, O_3.
- Un-adjusted average is biased: $\text{wt}(O_1, A_1, O_2, A_2, O_3) \neq \text{wt}(O_1, A_1, O_2, A_2, O_3, A_3)$.
- Adjusted weight: $\text{wt}(O_1, 1.5 \cdot A_1, O_2, 1.5 \cdot A_2, O_3)$.
- Adjusted average *may* be unbiased if $\mathbb{E}[\text{wt}(A_3)] = 0.5 \cdot \mathbb{E}[\text{wt}(A_1 + A_2)]$.
- But no reason assume this in general. In other words:

$$\mathbb{E}[\text{Outcome}|X, \text{Missing/unsampled}] \neq$$
$$\mathbb{E}[\text{Outcome}|X, \text{Not missing/sampled}]$$

Addressing non-response bias in your research

- Weighting (partial solution; see above).
- Use administrative data.
  - Administrative data (e.g. from the Census, Bonica's campaign finance data) do not suffer as much from non-response. The mechanism by which the data is recorded is often independent from the target population we might want to study.
  - But be aware of the *mechanism–non-response tradeoff*. Administrative data often contains less insight about mechanisms behind a relationship. We might see a change in individual-level contribution habits, but we do not necessarily know "why" this change occurred (was it due to changing attitudes about taxes, or due to evolving anxieties of job loss?)
- Consider incentive to encourage survey response.
  - But potential bias in who responds to incentive (self-selection).
  - MTurk survey participants often receive compensation for their participation in surveys, but are often quite familiar with standard treatments used in experiments (e.g. priming of race).

Social desirability bias & cheap talk

- *The Bradley effect - generalized*. The Bradley effect attempts to explain why public opinion polls often overstate support for non-white candidates in the US. Idea: voters who intend to vote for the white candidate tell pollsters they are undecided or are going to vote for the non-white candidate due to the social desirability of saying those things.
- This theory can easily be generalized to explain why polls may understate support for a referendum or candidate which is perceived as being socially objectionable. Research suggests that this effect is more pronounced for phone surveys (compared to online surveys).
- *Cheap talk*. No incentive for survey respondents to be truthful.

Addressing social desirability bias & cheap talk in your research

- Use administrative data on behavioral outcomes.
  1. Behavioral outcomes are often preferable to study because actors make more serious commitments towards realizing their goal. Social desirability issues will be minimized, since the bottom line of individuals will be affected by making one choice or another.
  2. Same problem as before. Behavioral data often is sparse in providing insight about causal mechanisms.
- Include the Marlowe Crowne scale in your survey. This scale is a battery of questions that assess whether a respondent has a tendency to give socially desirable responses.
  - Example items: "If I could get into a movie without paying and be sure I was not seen, I would probably do it." "I am sometimes irritated by people who ask favors of me."
  - Downweight respondents who show high social desirability according to this metric (but this is ad hoc).

Addressing social desirability bias & cheap talk in your research, continued

- Measure physiological responses images or primes.
  - These unconscious reactions are hard for participants to manipulate. Reactions include skin conductance (measure of arousal), heart rate, brain firing, eye movement, etc.
  - Measuring these responses can be expensive, and it is hard to access study participants in remote regions. This problem may be reduced with the growing ability of smart phones to measure physiological reactions.
  - See Petersen et al. 2015, who show that individuals who are physiologically aroused by party platforms much more likely to hold more partisan political views.

Insights for analysis

1. **Weighting is necessary but dangerous**.
2. **Errors in polls are correlated**.
3. **Historical trends are crucial to keep in view**.

Insights for analysis

1. Models upweight likely voters; this is a dangerous necessity.
   - *The "past is no guide" problem*. Who is actually a likely voter can be highly variable, especially for issues or candidates which have no clear precedent (e.g. Brexit). Historical turnout is not always predictive of future turnout (related to idea of concept drift).
     - Related to idea of concept drift in machine learning:

$$f_{X,Y,T=t}(x,y) \neq f_{X,Y,T=t+\mathrm{d}t}(x,y).$$

   - *Survey data on turnout have problems too*. Estimates of turnout based on surveys are susceptible to the problems discussed above (non-response, etc.).
   - **Solutions?**

Changes in data over time

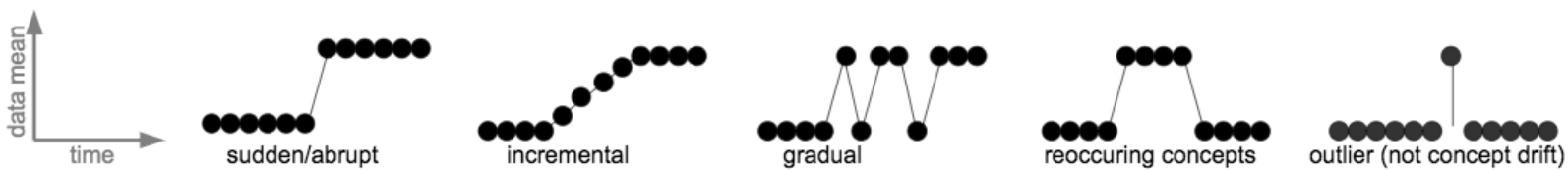
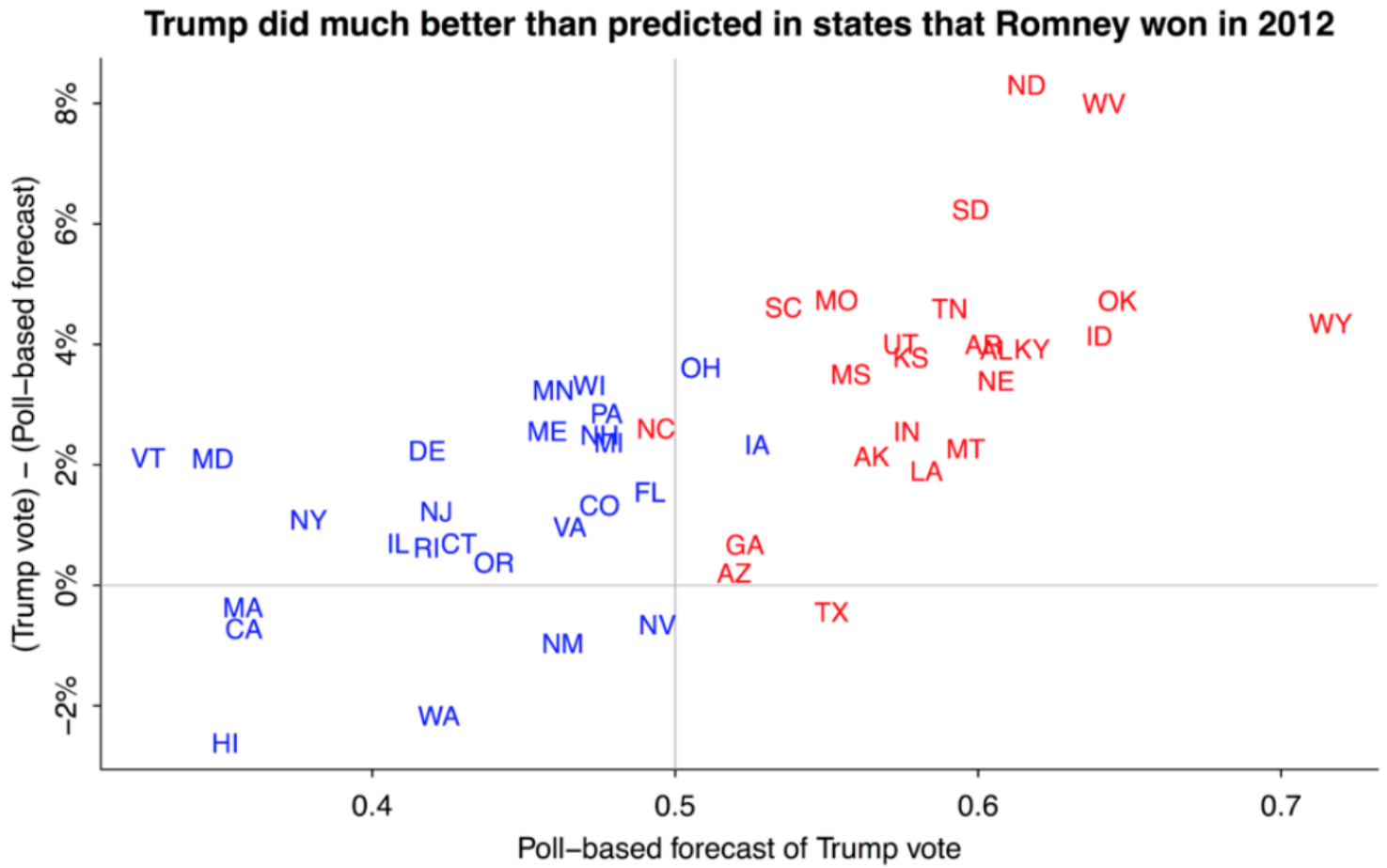


Fig. 2. Patterns of changes over time (outlier is not concept drift).

Insights for analysis, continued

2. Errors in the polls are correlated.
   - In a given election cycle, polling agencies tend to collectively over- or under-estimate support for controversial issues due to the issues just discussed. In other words, the poll residuals (actual vote − forecast) will be correlated across polling agencies.
     - *Thus, averaging polls does not reduce uncertainty as much as modelers often think.*
   - It is important to allow for high correlations in the residuals across the polls (and also across state, assuming the pollsters are more wrong in some regions than in others). Modelers should estimate results given different assumptions about these (unobserved) correlations as a sensitivity analysis.

Correlated residuals



This figure explains why the election forecasting in the 2016 presidential election was deeply problematic.

- On the one hand, the poll averages were off by about 2% nationally (Trump received 48% instead of 46% of the popular vote). This doesn't seem like much.
- On the other hand, the predicted returns were not over- and under-estimating Trumps support in a random fashion. Instead, there were severe correlations in the errors.
  - *Almost no where was Trump's support underestimated, and it was more severely underestimated in key regions than elsewhere*.
  - This is the key point.

Insights for analysis, continued

3. Historical trends are important - but watch out.
   - Forecast models that explicitly model historical context tend to perform better.
     - Examples of US-level contextual factors. Is the incumbent party candidate the sitting president? Is the economy in recession? Was the incumbent administration mared by social instability or foreign policy failures?
     - Yet, it is unclear how to weight these factors compared to current polling results in a quantitative model.
   - Moreover, even with 1000s of polling results, we don't have much information about over-time trends (only 12 elections in the US since 1972).
     - "Big data does not exist."

Discussion

- What else can we learn about quantitative political research from the recent polling breakdowns?
- What other methods for overcoming social desirability bias have you encountered?
- Is this discussion relevant to the "crisis in replication" in the social sciences?[1] If so, why? If not, why not?

[1] The crisis in replication refers to the ways in which many results in the social sciences cannot be replicated by other researchers

Take aways

- Any research consists of two key tasks: quantification (measurement) and analysis (often done statistically).
- In methods classes, we focus much of our effort on analysis. Yet, measurement is often more important. **Measurement trumps analysis**.
- Survey research is in a state of decline. It is time to think of more innovative ways to measure political attitudes.
- For example, we should consider using data from on behavioral outcomes or from physiological responses where possible. But be aware of the tradeoffs.

Acknowledgements

- These slides build from material presented by Pearl, Rubin, Jordan, Jackman, and others.