The Divided (But Not More Predictable) Electorate

A Machine Learning Analysis of Voting in American Presidential Elections

Seo-young Silvia Kim[1] and Jan Zilinsky[2]

April 16, 2021, MPSA 2021

[1]American University

[2]New York University

How predictable is the electorate?

Key quantity: *Predictability of vote choice.*

Research Question: Is differentiating between Republican and Democratic voters becoming easier?

How predictable is the electorate?

Key quantity: *Predictability of vote choice.*

Research Question: Is differentiating between Republican and Democratic voters becoming easier?

Result: With easily visible (race, gender) or discoverable (education, income, age) voter traits, inferring vote choice is as difficult today as half a century ago.

Strategy: Use hypothetical information sets.

Concepts

- Groups/coalitions: As a member of X, I will support Y.
- Partisanship: “I’m a strong Democrat,” “I’m an Independent,” $\cdots$
- Symbolic ideology: “I’m liberal,” “I’m conservative,” $\cdots$
- Operational ideology = issue positions: “I’m against abortion,”
  “I’m supportive of Medicare for All,” “I support COVID-19 lockdowns,” $\cdots$

Sorting by Groups

- Partisan/ideological sorting =
  convergence of symbolic/operational ideology with partisan identities (Levendusky 2009, Fiorina 2011)
- Social sorting =
  convergence of social identities and partisan identities e.g., race, religion, ... (Mason 2016 and 2018)

Why Is Group Sorting Important?

- Affective polarization and cross-cutting communication
- Group-level leverage in representation (“taken for granted”)
- Campaigns segment electorate into groups (perceptions)
  $\rightsquigarrow$ Practical implications
  If no swing voters, less effort in persuasion + more base mobilization
- Reasons to suspect increasing sorting
  e.g., 2016 Trump election, the diploma divide, white working-class men
- Popular claim: Partisanship is now a super-identity

Research Questions

Are demographic labels (increasingly) reliable predictor of vote choice?

- Hypothesis: if demographic sorting increases, the ability to infer vote choice based on demographics should also increase
  Ability to infer vote choice over time = intuitive measure of political alignment/sorting
- Membership in demographic groups $\rightsquigarrow$ social identity for many voters

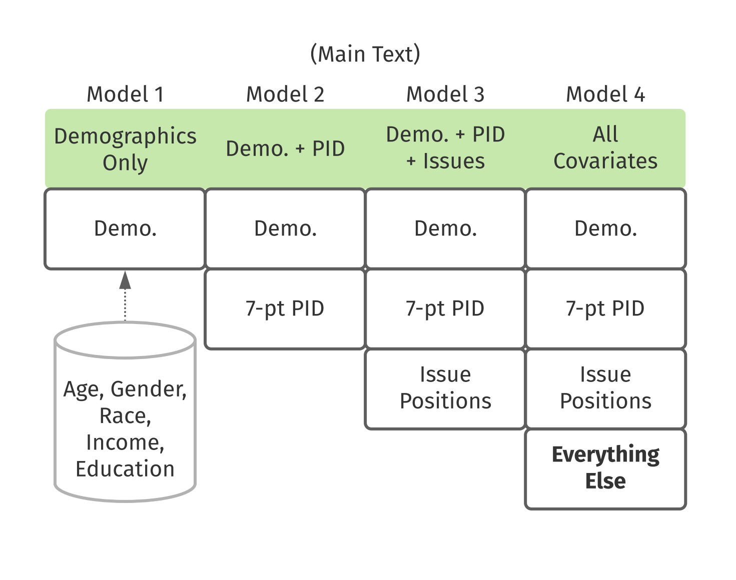
Operationalization and Hypotheses

- Demographic variables = race, education, income, age, gender
- Hypotheses
  1. *(Increasing Demographic Sorting)*: Vote choice will become increasingly predictable based on voters' demo. alone
  2. *(Increasing Party ID Sorting)*: Including explicit PID will make prediction increasingly easy over time + accuracy $\uparrow$
  3. *(Sufficiency of Party ID)*: Beyond PID and demo., other characteristics (e.g., issue positions) will contain minimal diagnostic information about vote choice

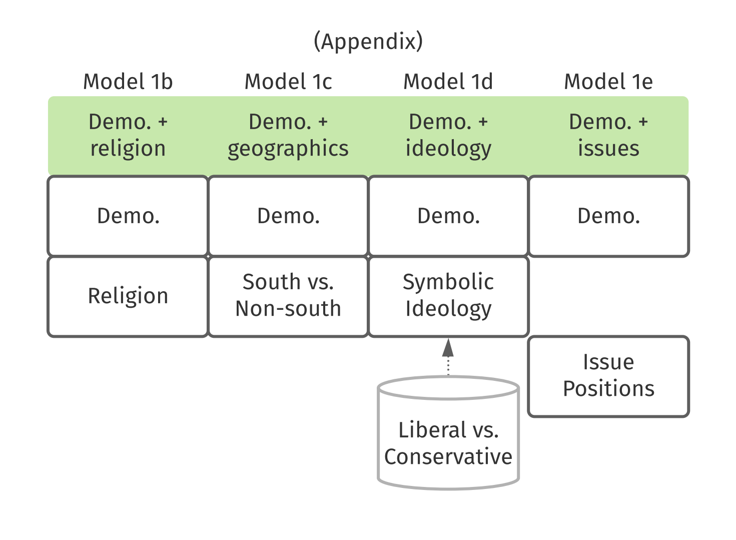
Data & Strategy

- Predict (out-of-sample) presidential vote choice with on the basis of a (potentially large) set of features
- Three national surveys:
  1952–2016 ANES, 2008–2018 CCES, 2020 Nationscape
- Prior research does not look into predictability

Main specifications



Additional specifications



Method: Tree-based Models (Supervised Machine Learning)

Random forests (Breiman, 2001)

- Performance-based on correct out-of-sample predictions (training/testing paradigm with cross validation, prevents overfitting)
- Flexible interaction structures possible
- High performance across a wide array of datasets

For an extensive review between prediction algorithms vs. traditional regressions, see Efron (2020)

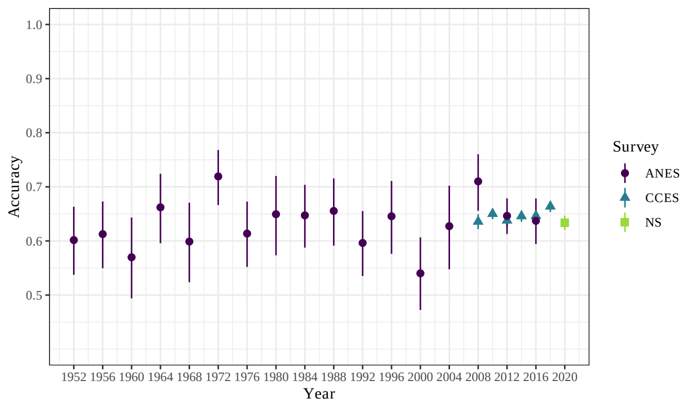
Classification Performance Metric

Definition of accuracy: proportion of correctly classified observations

	Actually Biden	Actually Trump
Expected Biden	180	50
Expected Trump	20	150

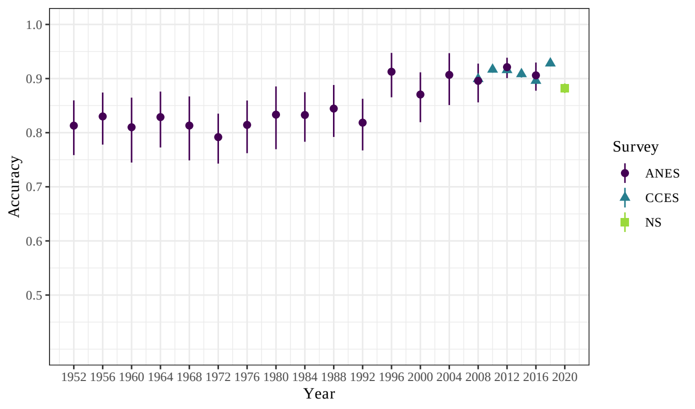
- Accuracy = (**TP** + **TN**) / (**TP** + **TN** + **FP** + **FN**) where
  - TP = true positive
  - TN = true negative
  - FP = false positive
  - FN = false negative
- In this example, (180 + 150) / (180 + 50 + 20 + 150)
- Also consider additional performance metrics: AUC, F-1 score

Results: Prediction Based Only on Demographics



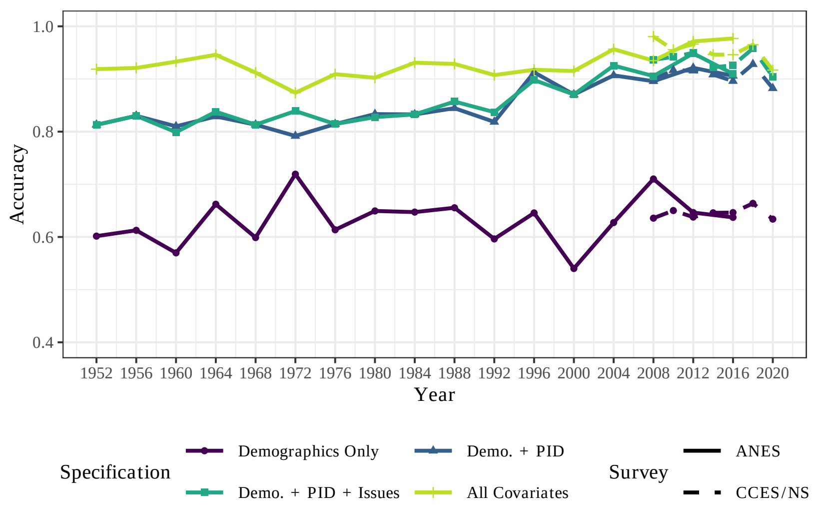
- Average accuracy across all surveys and waves is 63.5%. 63.1% for ANES, 64.7% for CCES, and 63.4% for Nationscape.
- Not increasing over time (regression slope p-value 0.24)

Results: Prediction Based on Demographics + 7-point Party ID



- Predictability increases when PID is included
- In line with other results on partisan polarization

Performance Metrics for All Four Models



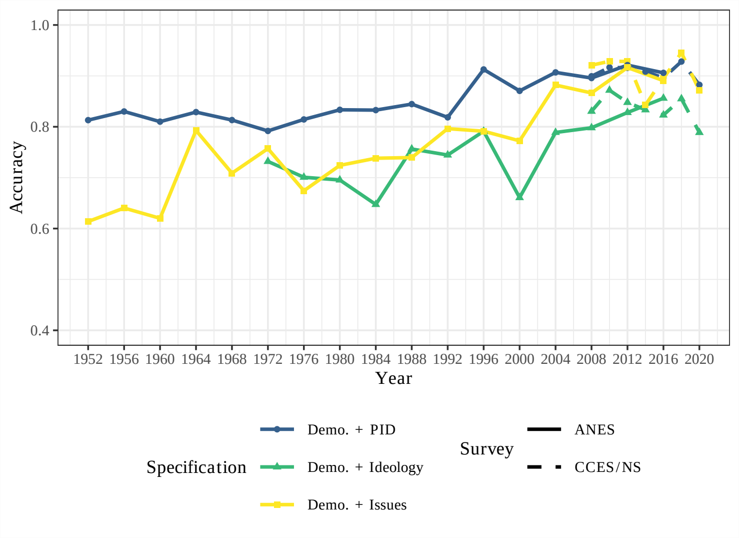
- Other covariates do contribute to increasing predictability
- Occupation, subjective class identification, group attitudes, beliefs, ...

Conclusion

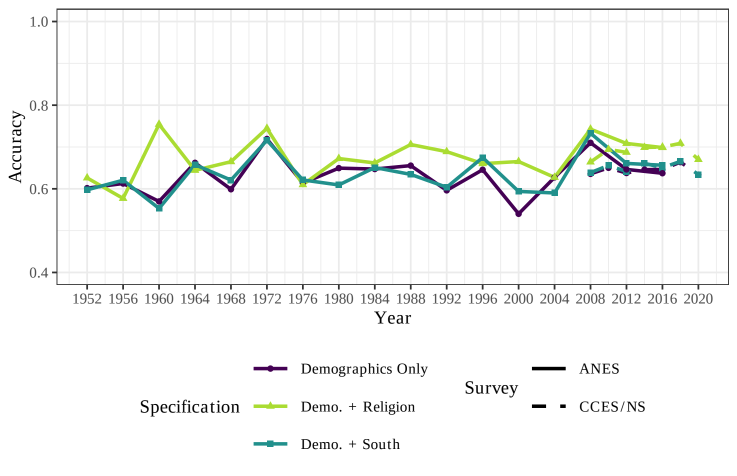
- Demographics function as important social groups
- **Using random forests, accuracy based on demographics-only is low and not increasing over time, while increasing for models 2–4**
- Predictability of vote choice is only 63.5%
- The electorate has not become more polarized along demographic lines in a way that is informative about voting behavior
- Operational ideology still matters (slightly)

Bonus Deck

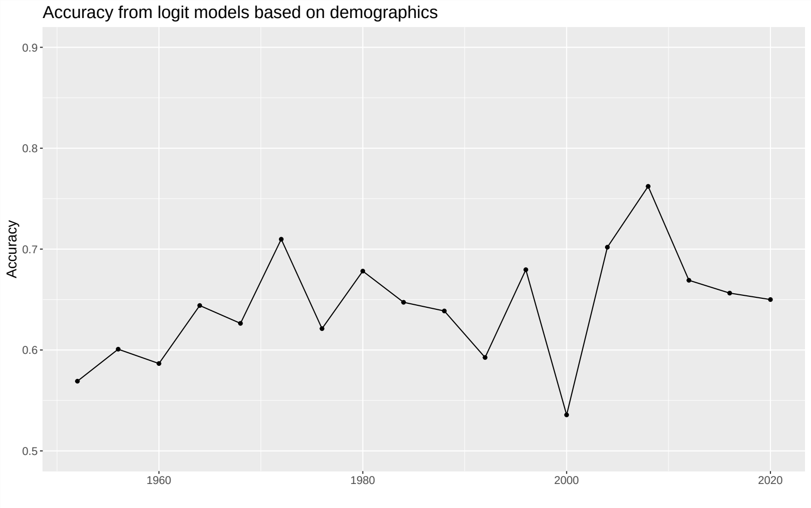
Additional models



Additional models



Logit (instead of RF)



P-value on the regression coefficient: 0.091.