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]

March 22, 2021

[1]American University

[2]New York University

Class / Education cleavage in voting behavior

Obama lost the white non-college vote by 10 p.p. in 2008 and by over 20 p.p. in 2012.

The diploma divide widened in 2016.

Class / Education cleavage in voting behavior

Obama lost the white non-college vote by 10 p.p. in 2008 and by over 20 p.p. in 2012.

The diploma divide widened in 2016.

Victory margin over Clinton among white non-college voters

National Exit Poll	Trump + 37
Pew Research	Trump + 36
ANES	Trump + 36
CCES	Trump + 24.3
Catalist estimates	Trump + 27.9
VOTER Survey	Trump + 22.5

CCES, VOTER, and exit polls: Own calculations.

For ANES and Catalist margins, see: https://medium.com/@yghitza_48326/what-happened-next-tuesday-e4e6637a4b81. Pew results: https://www.people-press.org/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/

Main RQ & Result

Main Q: 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.

Partisan Sorting by Groups

- Ideological sorting = Democrats are increasingly likely to be liberal and Republicans increasingly likely to be conservative
- 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

Is demographic sorting increasing? What proportion of voters are correctly classified with just demographic info?

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

Operationalization and Hypotheses

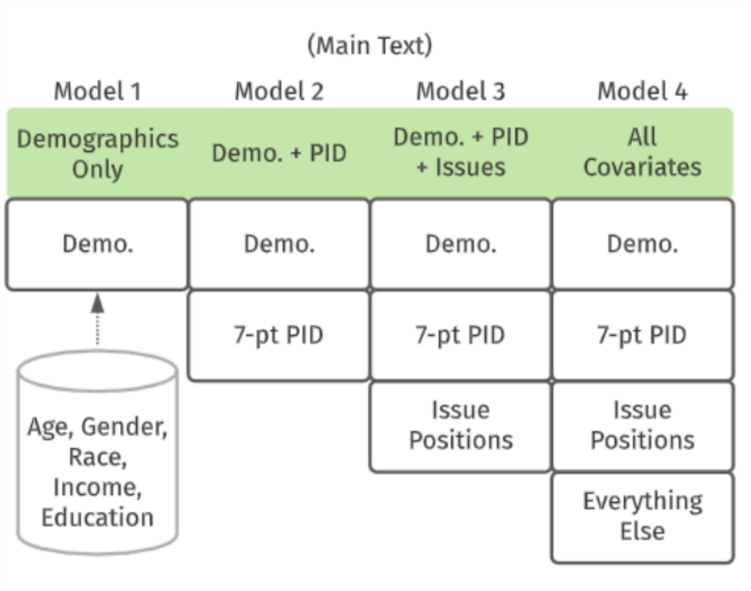
- Demographic variables = race, education, income, age, gender
- Hypotheses
  1. *(Increasing Demographic Sorting)*: Vote choice will become increasingly predictable based on voters' demo. alone
  2. *(Increasing Party ID Sorting)*: Including explicit PID will make predicting voting decisions increasingly easy over time, and accuracy will be higher relative to sparser models
  3. *(Sufficiency of Party ID)*: Beyond the initial sets of features (PID and demo.), other characteristics (e.g., issue positions) will contain minimal diagnostic information about vote choice

Data & Strategy

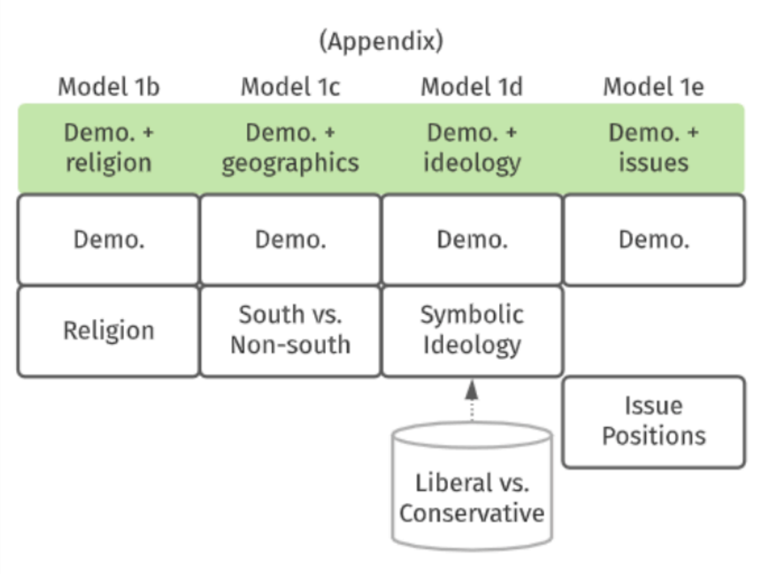
- Predict (out-of-sample) presidential vote choice with on the basis of a (potentially large) set of features
- Three national surveys:
  1952–2016 ANES, 2008–2018 CCES, 2020 Nationscape

- Prior research does not look into predictability
- **Using random forests, accuracy based on demographics-only is low and not increasing over time, while increasing for models 2–4**

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)

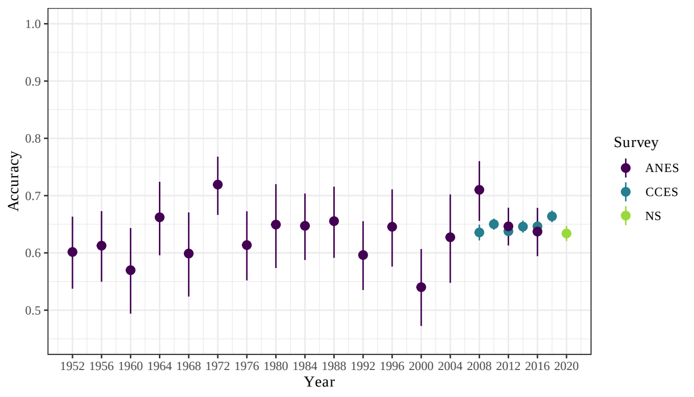
Classification Performance Metric

Definition of accuracy: proportion of correctly classified observations

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

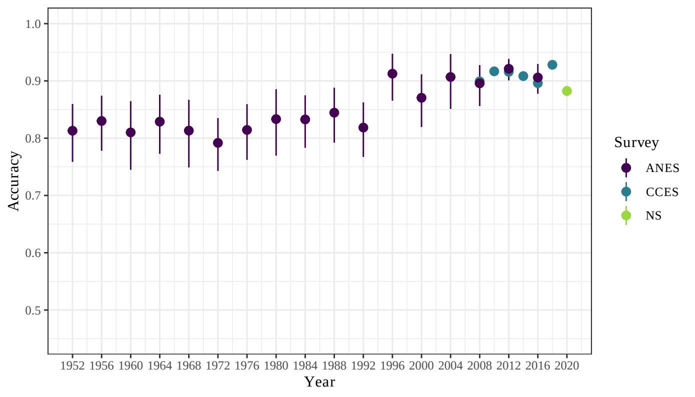
- Accuracy = (TP + TN) / (TP + TN + FP + FN) where
  - TP = true positive
  - TN = true negative
  - FP = false positive
  - FN = false negative
- In this example, (180 + 150) / (180 + 50 + 20 + 150)
- Also consider additional performance metrics: AUC, F-1 score

Results: Prediction Based Only on Demographics



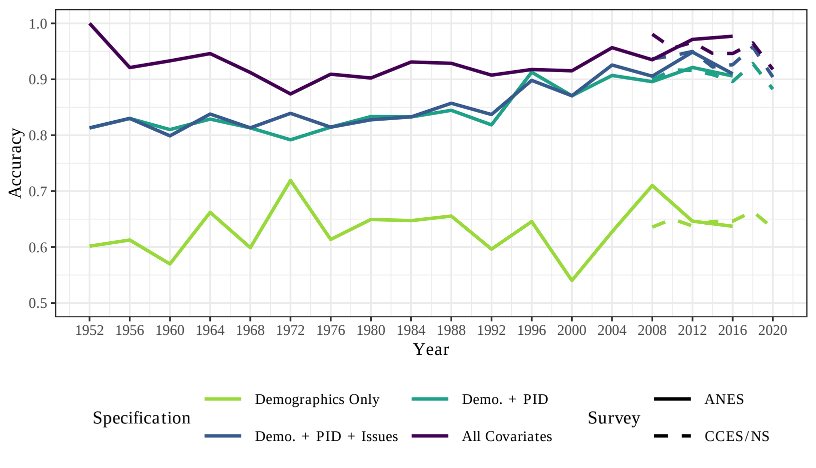
- Average accuracy across all surveys and waves is 63.5%. 63.1% for ANES, 64.7% for CCES, and 63.4% for Nationscape.
- Not increasing over time (regression slope *p*-value 0.24)

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, …

Variable Importance

Do demographics remain as top important variables after accounting for other variables?

Definition of permutation-based variable importance:

- Different from statistical significance
- Not variance explained
- How much does prediction accuracy decrease when a variable is randomly 'noised'?
- If removing/reshuffling variable greatly decreases accuracy, more 'important variable'

Variable Importance

Year	V1	V2
1952	Black	
1956	Income: 68-95 %tile	
1960	Age	
1964		
1968	Black	Age
1972	Black	
1976	Black	
1980	Black	
1984	Black	
1988	Black	
1992	Black	
1996		
2000		
2004		
2008	Black	
2012	Black	
2016	Black	

(a) PID/Issues Included (ANES)

Year	V1	V2
2018	Black	
2016	Black	
2014	Black	Age
2012	Black	
2010	Black	
2008		

(b) PID/Issues Included (CCES)

Demographics mostly disappears in S3. Identifying as Black = only consistent variable, but also disappears from top 10 in the full model.

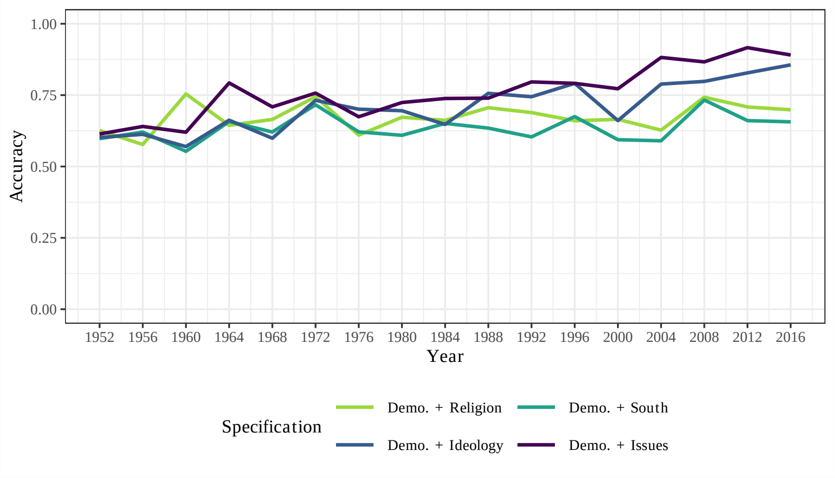
Conclusion

- Demographics function as important social groups
- To some degree, partisan sorting by demographics, but even with robust prediction model, not predictability for vote choice 63.5%
- In addition, demo. sorting not growing stronger over time
- Results validated from models with more covariates
- Demographics also generally not in permutation-based top 10 important variables in richer models

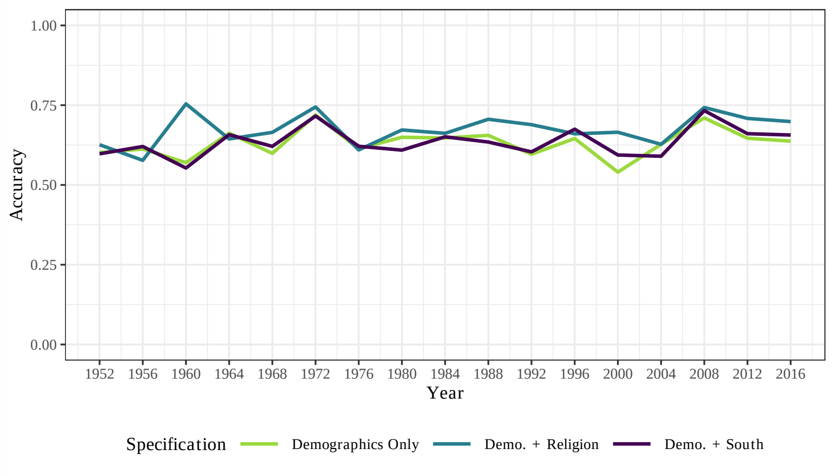
Electorate has not become more polarized along demographic lines a way that is informative about voting behavior

Bonus Deck

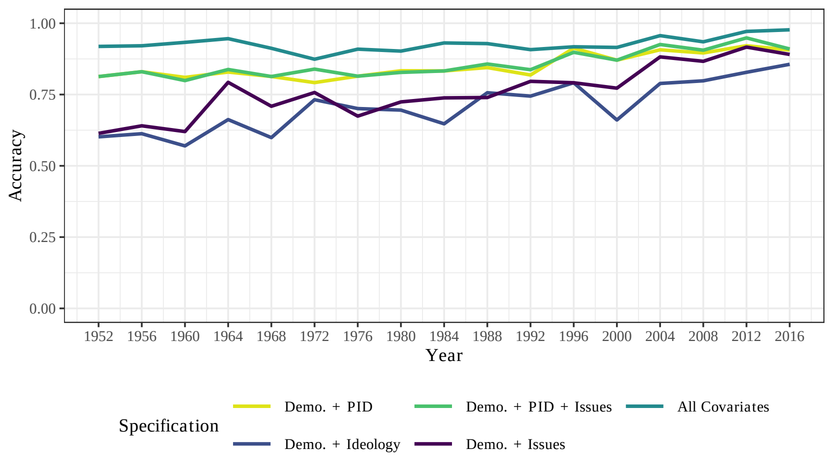
Additional models



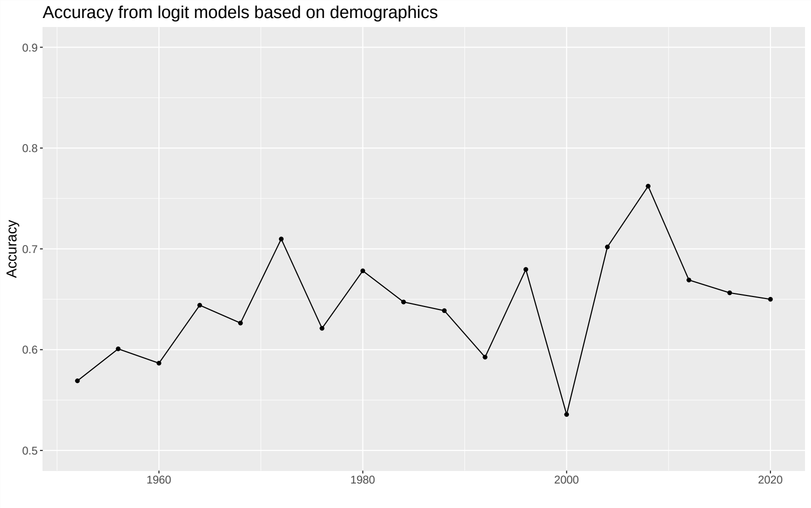
Additional models



Additional models



Additional models



P-value on the regression coefficient: 0.091.