30538 Final Project: Reproducible Research - Volunteerism, Engagement, and Polarization in the U.S.

Andrew White, Charles Huang, Justine Silverstein

2024-12-07

1. Background

This project began as a shared interest in trends behind volunteering rates in America, as two of our members (Justine and Charles) are AmeriCorps alumni.

For the past few years, concerns about the American public's increasing rates of isolation, decreasing lack of civic engagement and faith in institutions, and greater rates of political polarization have been prominent in the news and media. Our personal experiences with AmeriCorps and volunteering have taught us that volunteering can be effective at reducing isolation, increasing civic engagement/community awareness, and decreasing negative polarization towards "the other side". However, is volunteering a legitimate part of a public policy solution to these issues, or is it just a red herring?

Our research questions were: 1. What is the current state of volunteerism, political engagement and polarization in America? 2. What factors make people more likely to volunteer or be civically engaged?

2. Data Importing/Cleaning

Our datasets for this project were:

1. AmeriCorps CEV (Civic Engagement and Volunteering Supplement) for 2021
2. U.S. Census Bureau Volunteering and Civic Life Supplement - September 2021
3. ANES (American National Election Studies) Time Series Data, 2020

#1 and #2 primarily contain respondent information about volunteering and measures of civic engagement, while #3 contains information on political affiliation and polarization.

We are importing the data from the AmeriCorps and ANES websites. Because the datasets are over 100 MB, we include a Google Drive link here:

https://drive.google.com/drive/folders/1PUTN2pyh78MLoK0RVtGnf1ZwiM1BAAuV?usp=sharing

2a. Data cleaning - CEV/VCL data

As there are over 400 variables in the CEV and VCL data, here are the most relevant variables we focused on:

Frequency and Type of Volunteering: PES16: Did the respondent spend any time volunteering for any organization in the past 12 months? PES16D: Frequency of volunteering (e.g., basically every day, a few times a week). PTS16E: Approximate hours spent volunteering.

Political Engagement: PES2: How often the respondent discussed political, societal, or local issues with friends or family. PES5: How often these discussions occurred with neighbors. PES13: Contact or visits to a public official to express opinions. PES14: Boycotting or buying products based on political values or business practices.

Civic Participation and Group Membership: PES15: Belonging to groups, organizations, or associations in the past 12 months. Neighbor and Community Interaction: PES7: Participation in activities to improve their neighborhood or community. Voting Behavior: PES11: Whether the respondent voted in the last local elections.

Social Media and News Consumption: PES9: Posting views about political, societal, or local issues on the internet or social media. PES10: Frequency of consuming news related to political or societal issues.

Basic Demographics: Age: PRTAGE (Person's age) Gender: PESEX (Sex of the respondent) Race/Ethnicity: PTDTRACE (Detailed race and Hispanic origin) Marital Status: PEMARITL (Marital status of the respondent) Household Composition: HRNUMHOU (Number of persons in the household)

Potential Confounding Variables Income: HEFAMINC (Household family income level) Education: PEEDUCA (Highest level of school completed) Urban/Rural Status: GTMETSTA (Metropolitan or non-metropolitan status) Community Involvement: PES7 (Participation in neighborhood or community activities) Social Media Use: PES9 (Posting views about political, societal, or local issues on the internet or social media)

	hrhhid	hrmonth	hryear4	hurespli	hufinal	hetenure	hehousut	hetelhhd	hetelavl	hephoneo	...
0	333550513043249	9	2021	-1	226	2	1	1	-1	1	
1	900419145210736	9	2021	2	225	2	1	1	-1	1	

	hrhhid	hrmonth	hryear4	hurespli	hufinal	hetenure	hehousut	hetelhhd	hetelavl	hephoneo	...
2		206408571810641	9	2021	4	225	2	1	1	-1	1
3		606110727510599	9	2021	-1	225	1	1	-1	-1	1
4		848450067381002	9	2021	-1	216	-1	1	-1	-1	0

One data cleaning issue we encountered with the CEV/VCL data: the data is a mix of numeric code and qualitative input. We can create a mapping function to swap the numeric codes with qualitative input, but the existing qualitative input is outside of the data dictionary, so it won't get picked up by any mapping functions and will be transmuted into NaN data. We made a function that identifies all the values in the data that aren't picked up by our data dictionaries- this function is located in our config.py file.

For example, if there are some entries in a column that are already coded "Yes" or "No" in addition to "-1", "1", "2", etc. our existing mapping won't account for them and will turn them into NANs. We want to catch those and account for them.

CEV/VCL Data - Measuring Political Engagement

As part of our analysis, we gauged volunteerism by referring to the question "Did you volunteer in the last 12 months"? However, there isn't a single "civic/political engagement" question in the CEV/VCL data, but rather several different questions that are related. We chose five of the most relevant questions and weighted each based on their level of effort:

1. "How frequently do you talk to a family member/neighbor about politics?" (15%)
2. "How frequently do you post political views on social media?" (15%)
3. "How frequently do you consume political news/media?" (10%)
4. "Did you contact an elected official to express your opinion in the last 12 months?" (30%)
5. "Did you boycott a company based on their values in the last 12 months?" (30%)

This generated a score from 0 - 100 that we could use as a (imperfect) proxy for political engagement. We mutated a new variable, political_engagement_score, to measure this and added it to our dataset.

(Important caveat: For the political engagement questions, less than 20% of respondents answered three or more of the selected questions. To ensure meaningful data, we excluded all respondents who did not meet this threshold. While this approach improves the consistency of the dataset, we should be aware of potential selection bias.)

ANES data cleaning

For the ANES data, we create

```
Columns before adding engagement score:
['Household_ID', 'Household_ID_2', 'Volunteered_Past_Year', 'Volunteering_Frequency', 'Hours_

Shape before: (255744, 23)

Columns after adding engagement score:
['Household_ID', 'Household_ID_2', 'Volunteered_Past_Year', 'Volunteering_Frequency', 'Hours_

Shape after: (255744, 25)

Verifying engagement score columns exist:
'political_engagement_score' exists: True
'engagement_level' exists: True



Columns in saved CSV:
['Household_ID', 'Household_ID_2', 'Volunteered_Past_Year', 'Volunteering_Frequency', 'Hours_
Dataset saved to: shiny-app/basic-app/data/cev_2021_cleaned.csv
```

Data Cleaning - ANES Data

As with the CEV/VCL data, our goal was to subset the data so that it only contains relevant variables. We accomplish this by making two lists:

List 1: This list is designed to capture variables covering geographic information (V201011, V201013a, V201013b, V201014a, V201014b)

List 2: This list is designed to capture variables covering information about assessments of political positioning (i.e. left, right, center)

ANES - More Data Cleaning

Analyzing Question V201200, which is a question asking:

"Where would you place yourself on this scale, or haven't you thought much about this? Value Labels-9. Refused -8. Don't know 1. Extremely Liberal 2. Liberal 3. Slightly Liberal

4. Moderate; middle of the road 5. Slightly Conservative 6. Conservative 7. Extremely Conservative 99. Haven't thought much about this"

We use this data to make a dataframe aggregated by state, and then we can show correlation between measure of polarity and the share of respondents in a state who did volunteer work.

Note that we previously used two ANES variables as US State variables, with only "US State 2" being used in analysis, purely because it has more in-universe entries. This appears to be partially due to respondent reactions to different questions, and partially due to information restrictions on the dataset. As such, "US State" is only used for CEV data.

Then, make data using V201228, which asks:

"Generally speaking, do you usually think of yourself as [a Democrat, a Republican / a Republican, a Democrat], an independent, or what?"

-9. Refused -8. Don't know -4. Technical error 0. No preference {VOL - video/phone only} 1. Democrat 2. Republican 3. Independent 5. Other party {SPECIFY}

2. Exploratory Analysis

2a. Measures of Polarization

We will use two measures of polarization from the ANES data; each provides some amount of information that can be interpreted to indicate polarization to a certain extent, though both have their drawbacks.

1. Share of Outliers - we create a series of functions that group respondents by party Democrats with conservative-leaning ideologies, and Republicans with liberal-leaning ones.

	Party_Affiliation_(V201228)	Outliers	Party_Count	Percent_Outliers
0	Democrat	0	137	0.0000
1	Independent	0	0	0.0000
2	Other party	0	0	0.0000
3	Refused	0	0	0.0000
4	Republican	4	84	0.0181

In a paper on quantifying polarization written by Aaron Bramson et al (https://inferenceproject.yale.edu/sites/de the authors examine a range of polarization indicators. A relatively simple (and in some ways problematic) measurement is called spread, or dispersion- essentially the gap between the most extreme political positions.

In the paper, Bramson et al. explain: “Polarization in the sense of spread can be measured as the value of the agent with the highest belief value minus the value of the agent with the lowest belief value (sometimes called the ‘range’ of the data).” ”

We (imperfectly) approximate this using two more variables: V201206 and V201207. These ask respondents to position political parties on the political spectrum. We can select the most ideologically distant nodes on the personal ideology scale (extremely liberal and extremely conservative) and capture how far apart their conceptions of each party are, on average, and then disaggregate by state.

We will assign the different ideological positions to different points on a spectrum, namely: -3, -2, and -1 are “Extremely Liberal”, “Liberal”, and “Slightly Liberal”;

and: 0 is “Moderate; middle of the road”;

finally, 1, 2, and 3 are “Slightly conservative”, “Conservative” and “Extremely conservative.” ”

We’ll then compare average positions by state. For example, if the average extremely liberal respondent in Texas places Democrats at at -1 (slightly liberal) and the average for the extremely conservative respondents in -3 (extremely liberal), then the distance between the two is 4, meaning Texas would have a spread of 4 for this question.

We create the crosswalks V201206 and V201207 -

We use the Positioning columns to compare the average party position selections between the 2 extremes.

Step 1: First grouping

Step 2: We use the variable position_groups to create a dataframe that has 4 columns: (1) The position Liberals give Democrats on the spectrum (2) the position Conservatives give Democrats on the spectrum (3) the position Liberals give Republicans on the spectrum (4) the position Conservatives give Republicans on the spectrum

Step 3: Now, we use those 4 columns to create the spread, meaning the absolute value of the difference betweeen:

(1) The position Liberals give Democrats on the spectrum and the position Conservatives give Democrats on the spectrum

(2) The position Liberals give Republicans on the spectrum and the position Conservatives give Republicans on the spectrum

Step 4: Now, we graph those differences and interpret these spreads as an indicator of polarization. We acknowledge that this is too simple an analysis to account for the full complexity of this kind of measurement, as polarization involves not just distance between extremes, but also clustering around them. We also acknowledge that our political scale assumes a linear ideological spectrum, which isn’t always the case.