Datasheet for '2022 Cooperative Election Study (CES) Common Content dataset'*

Jinyan Wei

2 December 2024

This datasheet provides a comprehensive overview of the 2022 Cooperative Election Study (CES) Common Content dataset. The dataset includes responses from a nationally representative sample of 60,000 American adults, offering insights into voter behavior, demographic trends, and policy preferences. The dataset features key variables such as vote choice, demographic characteristics, income, and education levels, making it a valuable resource for studying electoral dynamics in the United States. This final release includes survey data, vote validation conducted by TargetSmart, and accompanying documentation for detailed exploration.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The 2022 Cooperative Election Study (CES) dataset was created to study American voters' electoral behavior, political opinions, and experiences, with a focus on understanding representation and accountability during elections. It specifically aims to provide large-scale, representative data on voter preferences and behaviors across legislative constituencies, addressing the need for comprehensive datasets that capture both state and national dynamics.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The CES dataset is a collaborative effort led by principal investigators Stephen Ansolabehere (Harvard University), Brian Schaffner (Tufts University), and Marissa Shih (YouGov), involving 60 research teams and organizations. It is conducted under the broader framework of the Cooperative Congressional Election Study, which was renamed the Cooperative Election Study in 2020.

[*] Code and data are available at: https://github.com/jeno0403/Voter-Behavior-2022-CES.git.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

   - The creation of the dataset was supported by the National Science Foundation under Award #2148907. Additional research support was provided by Harvard University, Tufts University, and other participating universities and organizations.

4. *Any other comments?*

   - The CES dataset is part of an ongoing series of election studies initiated in 2006. It employs advanced sampling and weighting techniques to ensure a high level of accuracy and representativeness. The dataset is a valuable tool for scholars, policymakers, and analysts, offering detailed insights into voter behavior and public opinion trends across diverse demographic groups and political contexts.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

   - The dataset comprises survey responses from U.S. residents regarding their voting behavior, political opinions, demographic characteristics, and contextual information about electoral districts. The dataset includes multiple types of instances, such as individual respondents, their answers to survey questions, and validated voter records. Relationships between instances, such as geographical or demographic correlations, are also embedded within the data.

2. *How many instances are there in total (of each type, if appropriate)?*

   - The dataset contains responses from approximately 60,000 individuals, divided between pre-election and post-election waves, representing one of the largest sample sizes for such studies.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset is a sample drawn using matched sampling techniques to ensure representativeness of the U.S. adult population. It aligns with demographic and voter registration distributions based on sources like the American Community Survey and voter files. Geographic and demographic balancing was validated through weighting and comparisons to actual election results.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of processed survey responses, including demographic variables (e.g., age, gender, race, education), political preferences (e.g., party affiliation, policy support), and voting behavior (e.g., voter turnout, candidate selection). Contextual data about electoral districts and representatives are also included.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Some instances include targets, such as validated voter turnout or party preference, which can be used for predictive modeling and analysis.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - Some instances may have missing responses due to participant non-response or skipped questions. For validated voter records, unmatched respondents may lack certain data due to incomplete or inaccurate registration information.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Relationships between instances, such as demographic clustering, geographical distribution, and voting trends, are captured through variables like state, congressional district, and demographic attributes.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - The dataset is often divided into pre-election and post-election waves, and weights (e.g., commonweight, vvweight) are provided for analyses targeting different populations (e.g., registered voters, validated voters).

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - Errors may arise from self-reported data (e.g., misreporting) and sample matching (e.g., unmatched voter records). These are mitigated through validation and weighting techniques.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there*

official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

- The dataset relies on external voter file data for validation (e.g., TargetSmart), which is integrated into the dataset with weights and additional variables. Archival versions are maintained, but external dependencies may evolve over time.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The dataset does not contain legally confidential data but includes anonymized voter registration and demographic details to preserve respondent privacy.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - Survey questions and responses may touch on sensitive topics, such as political opinions, voting rights, and demographic factors, which could evoke strong reactions or discomfort among some audiences.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Sub-populations are identified through demographic variables such as age, gender, race, and political affiliation. Distributions are adjusted to ensure representativeness of these groups within the overall sample.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - The dataset is anonymized to prevent direct identification of individuals. However, indirect identification risks exist if combined with other datasets containing overlapping variables.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The dataset includes sensitive information, such as political opinions, party affiliation, race, ethnicity, and demographic variables. These are anonymized to ensure data security.

16. *Any other comments?*

    - The CES dataset provides a rich source for understanding voter behavior and preferences, with built-in safeguards for confidentiality and representativeness. It supports both exploratory analyses and targeted research on U.S. electoral dynamics. Further information can be found in the CES FAQs.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was collected through surveys conducted by YouGov using its matched random sample methodology. Respondents were recruited from opt-in internet panels, and their survey responses were directly reported. The data was validated by matching to TargetSmart's voter database to verify voter registration and turnout records.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Surveys were administered online using YouGov's platform. Respondents were matched to a target sample based on demographic, geographic, and voter data variables derived from high-quality datasets, including the American Community Survey (ACS) and voter registration files. Matching was validated using weighted Euclidean distance metrics to ensure representativeness.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - A stratified sampling approach was used. The sample was matched to a politically representative modeled frame based on variables such as age, gender, race, and education. Weighting adjusted for imbalances to ensure representativeness of U.S. adults and registered voters.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Respondents were recruited from opt-in panels provided by YouGov and other partner organizations. Compensation details were managed by YouGov but typically included incentives customary for panel participation, such as monetary rewards or points.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - Data collection occurred in two waves: pre-election (September 29 – November 8, 2022) and post-election (November 10 – December 15, 2022). Validation was completed in August 2023.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - All respondents provided explicit consent to participate in the study. While specific institutional review board (IRB) details are not outlined, CES adhered to industry standards for ethical survey research, including anonymization of responses.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - Data was collected directly from respondents via online surveys and supplemented by third-party voter records from TargetSmart for validation purposes.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - Participants were notified about the study's purpose during recruitment and consented to participate before completing the survey.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - Yes, the individuals in question consented to the collection and use of their data.Respondents provided their consent by agreeing to participate in the study at the start of the survey. Those who did not consent were not included in the dataset.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Respondents could choose not to answer specific questions or discontinue participation at any time.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - The CES data is anonymized and used strictly for research purposes, minimizing potential adverse impacts on data subjects. Specific data protection impact analyses are not detailed in the guide.

12. *Any other comments?*

    - The dataset is anonymized, ensuring respondent privacy and adherence to ethical research practices. All data collection complied with best practices for online surveys.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, the CES 2022 dataset underwent several preprocessing and cleaning steps to ensure high data quality. These included the removal of incomplete or low-quality responses, such as those where respondents did not complete critical sections or provided inconsistent answers. Survey data was matched to TargetSmart voter files to validate voter registration and turnout information. Weighting and adjustment procedures were applied to correct for sampling imbalances and to ensure the dataset's representativeness.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - Yes, the raw data has been saved and is accessible through the Harvard Dataverse repository. Researchers can access this raw data to conduct unanticipated analyses or validate preprocessing steps. Access to the raw data requires registration and adherence to data use policies. The dataset can be found here.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Preprocessing and cleaning were conducted using YouGov's proprietary tools and widely-used statistical software, such as R and Python, for data processing, weighting, and analysis. Specific code for academic modules may also be available upon request.

4. *Any other comments?*

- The dataset provides both cleaned data for ease of use and raw data for flexibility in research. The preprocessing steps ensure that the data is ready for analysis while maintaining transparency about modifications.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - The CES dataset has been extensively used for tasks such as analyzing voter turnout, partisan polarization, issue-based preferences, and demographic influences on voting patterns. It has also been utilized in studies exploring congressional representation, campaign effectiveness, and public opinion trends.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - Yes, a repository of papers and research that use the CES dataset is available on the CES website and the Harvard Dataverse. These repositories provide access to publications and systems that leverage the dataset. For more information, visit CES Publications.
3. *What (other) tasks could the dataset be used for?*
   - The dataset can be used for a wide range of tasks, including:
     1. Predictive modeling of election outcomes.
     2. Analysis of voter demographic trends and geographic variations.
     3. Studying the relationship between political preferences and social or economic variables.
     4. Examining voter behavior in response to political messaging or events.
     5. Conducting longitudinal studies on changes in public opinion.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - The CES dataset uses weighted and matched samples to ensure representativeness, but researchers should be cautious about interpreting results for small subpopulations, as these may have higher variability. The self-reported nature of survey responses may also introduce biases, particularly for sensitive questions. Researchers

should be aware of these limitations to avoid unfair treatment of individuals or groups. To mitigate risks, pairing analysis with validation steps and contextual understanding is recommended.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - The dataset should not be used for commercial purposes, political campaigning, or any activities that could violate respondent privacy or anonymity. Additionally, it should not be used to target or stereotype individuals or groups based on sensitive attributes, such as race, ethnicity, or political affiliation.

6. *Any other comments?*

   - The CES dataset is a vital resource for understanding U.S. electoral dynamics, but users should adhere to ethical guidelines and data use policies to maximize its positive impact and mitigate risks. Let me know if you'd like this formatted further or expanded!

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - Yes, the CES dataset is publicly available for academic and research purposes. It is distributed through the Harvard Dataverse, which allows third parties such as researchers, policymakers, and students to access the data.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The dataset is distributed through the Harvard Dataverse as downloadable files. It has a digital object identifier (DOI) to facilitate citation and tracking. The DOI for the CES 2022 dataset is available through the Dataverse repository.

3. *When will the dataset be distributed?*

   - The dataset is typically distributed after data validation and cleaning are completed. For CES 2022, the validated dataset was made available in 2023.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - Yes, the dataset is distributed under a data use agreement (DUA). Users must agree to these terms before accessing the data, which specify restrictions on its use, such as non-commercial purposes only. More details are available on the CES website and Dataverse.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - Yes, some restrictions apply due to the integration of third-party voter file data, such as data from TargetSmart. Users are required to comply with additional terms outlined in the data use agreement.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No explicit export controls or regulatory restrictions are mentioned for the CES dataset. However, users are expected to comply with all relevant laws when accessing and using the data.

7. *Any other comments?*

    - The dataset is an invaluable resource for understanding voter behavior and public opinion in the U.S. Its distribution model ensures accessibility while maintaining ethical and legal standards.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset is hosted and maintained by the Cooperative Election Study team and supported by the Harvard Dataverse infrastructure.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - Inquiries about the dataset can be directed to the CES team via the contact information provided on their website.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - If any issues or errors are identified in the dataset, they are addressed through updates or errata posted on the Dataverse page or communicated to users via the CES website.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- Yes, the dataset may be updated to correct errors or include additional modules. Updates are communicated through the Dataverse repository and CES mailing lists.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - The dataset is anonymized, and no personal identifiers are retained, ensuring compliance with data protection standards. Retention limits do not apply to anonymized data.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - Older versions are archived and remain accessible on the Harvard Dataverse to ensure reproducibility of past research.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - Extensions or contributions are not directly incorporated into the dataset but may be shared through publications or independent repositories. Validation and communication of these contributions are the responsibility of individual contributors.

8. *Any other comments?*

   - The CES team ensures ongoing support and accessibility for the dataset while maintaining high ethical standards. Researchers are encouraged to engage with the data responsibly and contribute to its growing body of related research.

1 References