Codebook "Do More Disaggregated Electoral Results Deter Aggregation Fraud?" (BJPS)

Miguel R. Rueda* Guy Grossman† Shuning Ge‡

June 9, 2025

Dataset #1: mainData_period.csv

This is the main dataset that is used in the analysis. The unit of observation is country - election-type - period.

- **country_name**: The country name.
- **country_election**: A factor variable with country-election-type identifier.
- **election_type**: The election type: Legislative or Presidential.
- **cowcode**: Country code.
- **period_num**: The period number: 1 = year 2000 to 2004, 2 = year 2005 to year 2009, 3 = year 2010 to year 2014, 4 = year 2015 to year 2020.
- **period_char**: A factor variable created from period_num: period0005 = year 2000 to 2004, period0509 = year 2005 to year 2009, period1015 = year 2010 to year 2014, period1520 = year 2015 to year 2020.
- **log_nlevel**: Logged number of the smallest administrative units at which electoral results are published.
- **policy_down_level**: "Granularity (level)." Dichotomous variable that takes the value of one for all periods in which the level of the most disaggregated published results is a more granular level than that of the first period. If the most disaggregated level

*Associate Professor, Department of Political Science, Emory University. Email: miguel.rueda@emory.edu. Corresponding author.

†David M. Knott Professor of Global Politics, Department of Political Science, University of Pennsylvania. Email: ggros@upenn.edu.

‡PhD student, Massachusetts Institute of Technology. Email: shuningg@mit.edu.

was used in the first period, all other periods will take the value of one unless the granularity level falls.

- **policy_down_u_std**: "Granularity (std)." Dichotomous variable that takes the value of one in periods where the number units at which electoral results are published is one standard deviation above the number of units in the first period. If the most disaggregated level was used in the first period, all other periods will take the value of one unless the number of the most disaggregated units falls by more than a standard deviation. The standard deviation is computed using the aggregation units of that country-election type.

- **policy_down_u_50p**: "Granularity (median)." Dichotomous variable that takes the value of one in periods where the number of units at which electoral results are published has grown by more than the median growth in the sample. If the most disaggregated level was used in the first period, all other periods will take the value of one unless the number of the most disaggregated units falls by more than the median decrease in the sample. Growth is defined with respect to the first-period number of units for that country-election type.

- **policy_down_u_g2**: "Granularity (moving average)." Dichotomous variable that takes the value of one in periods where the number of units at which electoral results are published has grown by more than the median growth in the sample. If the most disaggregated level was used in the first period, all other periods will take the value of one unless the number of the most disaggregated units falls by more than the median decrease in the sample. Growth is defined with respect to the previous moving average of the number of units for that country-election type.

- **other_irreg**: "Other voting irregularities." Taken from V-Dem variable *v2elirreg*. The original question asks if there was "evidence of other intentional irregularities by incumbent, opposition parties, and/or vote fraud" (V-Dem Codebook v11.1 p66). The original variable takes on values from 0 (Yes, there were systematic and almost nationwide other irregularities) to 4 (None, there was no evidence of intentional other irregularities). We multiplied the original variable by -1 so that the higher the values, the worse the voting irregularities are.

- **pei_faircount**: "Unfair count." Taken from PEI variable *faircount*. The question asks if votes were counted fairly with answers ranging from 1 (strongly disagree) to 5 (strongly agree). We multiply the original *faircount* by -1 so that the higher the value, the more unfair the vote counting is.

- **registry_irreg**: "Registry irregularities." Taken from V-Dem *v2elrgstry*. The question asks, "In this national election, was there a reasonably accurate voter registry in place and was it used" (V-Dem Codebook v11.1 p65)? The answers range from 0 to 4 where the higher the value, the more accurate the registry is. We again reversed the original values by multiplying -1.

- **pei_delay**: "Delay in announcement." Taken from PEI variable *delay*. The question asks whether the results were announced without undue delay. The answer ranges from 1 (strongly disagree) to 5 (strongly agree). We multiplied the original values with -1 so that the higher the values, the more undue delay the results announcement had.

- **govt_intimidation**: "Intimidation to opposition." Taken from V-Dem variable *v2elintim*. The original question asks, "In this national election, were opposition candidates/parties/campaign workers subjected to repression, intimidation, violence, or harassment by the government, the ruling party, or their agents?" (V-Dem Codebook v11.1 p67). The answers range from 0 ("Yes. The repression and intimidation by the government or its agents was so strong that the entire period was quiet.") to 4 ("None. There was no harassment or intimidation of opposition by the government or its agents, during the election campaign period and polling day."). We multiplied the original values by -1 so that the higher the value, the more intimidation.

- **nelda15**: "Intimidation of opposition." Taken from NELDA variable 15. This is an alternative measure of government intimidation and harassment of the opposition. A yes is coded 1, and No is 0.

- **nelda33**: "Severe violence, civilians." Taken from NELDA variable #33. This variable is coded 1 if there was significant violence resulting in civilian deaths at any point of the election (before, during, or after). 0 otherwise.

- **v2xnp_client**: "Clientelism." Taken from V-Dem variable *v2xnp_client*. Answers to this question measure how much politics rely on clientelism, ranging from 0 (low) to 1 (high).

- **elec_quality**: "Election fraud." Taken from V-Dem variable *v2elfrfair*. The question is, "Taking all aspects of the pre-election period, election day, and the post-election process into account, would you consider this national election to be free and fair?" (V-Dem Codebook v11.1 p70)? Similarly, the original answers range from 0 (the election were fundamentally flawed) to 4 (Yes, the election could be considered free and fair). We multiplying original values by -1.

- **fraud_monitored**: "Monitors-reported fraud." Taken directly from V-Dem variable *v2elwestmon*, but the source of this variable is NELDA variable 47. This variable is coded 1 if Western monitors reported allegations of significant vote fraud. 0 otherwise.

- **nelda29**: "Riots and protests." Taken from NELDA variable 29. This question asks if there were riots and protests after the election.

- **ecav_violence**: "Election violence." Taken from ECAV variable *eventViolence*. This variable takes teh value of 1 if the event is violent and 0 otherwise. The outcome variable election violence is calculated as the number of violent events for each country - election.

- **pre_violence**: "Pre-election violence." Coded based on *eventViolence* ECAV's violence variable and on *Date*, which is the event date. For more details, see below.

- **post_violence**: "Post-election violence." By comparing the event date with the election date, we separately counted the number of violent versus non-violent events before or after the election. Again, we constructed the pre-/post-election violence outcome variables as the count of pre-/post-election violence events. Violence events happened on the election date were counted as pre-election violence events.

- **freefair**: "Public Opinion." Taken from Barometers survey datasets. For details of harmonization and construction of the variable see the Supplementary Information.

- **opposition_prevented_nelda13**: "Opposition not allowed." Taken from NELDA variable 13. This question asks whether opposition leaders were prevented from running. We recoded the variable such that 0 represents no and 1 yes.

- **campaign_media**: "Free media." Taken from V-Dem variable *v2elfrcamp*. The question asks if "parties or candidates receive either free or publicly financed access to national broadcast media" in the national election. The answers include 0 ("either no parties or only the governing party receives free access"), 1 (some parties other than the governing party), 2 (all parties have access to free media). We multiplied the original variable by -1 so that the higher the value, the less free media access.

- **domestic_monitors**: "Domestic monitors." Taken from V-Dem variable *v2eldommon*. The original question asks, "In this national election, were election monitors from all parties and independent domestic election monitors allowed to monitor the vote at polling stations across the country?" (V-Dem Codebook v11.1 p69). We multiplied the original answers by -1 so that 0 indicates no and -1 indicates yes.

- **inter_monitors**: "International monitors." Taken from V-Dem variable *v2elintmon* asking if international election monitors were present in the national election. We multiplied the original answers by -1. 0 indicates no/unclear and -1 represents yes.

- **pei_laws**: "Electoral laws." Taken from PEI index *laws*. This index summarizes three compositing variables: *lawsunfair2* (if "election laws were unfair to smaller parties"), *favoredincumbent2* (if "election laws favored the governing party or parties"), and *citizens2* (if "election laws restricted citizens' rights"). The scale of this index is standardized to 100 point scale and the higher the values, the higher integrity. We multiply the original variable by -1 so that the higher the value, the less integrity.

- **pei_procedures**: "Electoral procedures." Taken from PEI index *procedures*. This index summarizes four compositing variables: *managed* ("elections were well managed"), *votinginfo* ("information about voting procedures was widely available"), *fairofficials* ("election officials were fair"), and *legalelections* ("elections were conducted in accordance with the law"). This variable has a 100-point scale. We multiplied this variable with -1 to make the higher value more negative.

- **incumbent_vote_share**: "Incumbent vote share." Manually coded using Wikipedia as the main data source, Psephos: Adam Carr's Electoral Archive[1] as well as CLEA as supporting data sources. For details on the coding criteria, refer to the Supplementary Information.
- **l_le_wb_pop**: "Logged total population." Taken from United Nations World Population Prospects variable *Estimates*.[2]
- **lv2elembaut**: "EMB autonomy." Taken from V-Dem variable *v2elembaut*.
- **lv2elembcap**: "EMB capacity." Taken from V-Dem variable *v2elembcap*.
- **le_polity2**: "Polity score." Taken from V-Dem variable *epolity2*.
- **l_lgdp_c15**: "Logged GDP." Taken from the World Bank variable *GDP (constant 2015 US$)*.
- **lurban**: "Urbanization." Taken from the World Bank variable *Urban population (% of total population)*.
- **ltotalseats**: "Size of the legislature." Taken from The Database of Political Institutions (DPI) variable *totalseats*.
- **lnelda45**: " International monitors." Taken from NELDA variable *nelda45*.
- **nelda24**: "Losing dummy." Taken from NELDA variable 24. This variable is coded 1 "if the party associated with the incumbent lost", 0 otherwise.
- **le_wb_pop**: The variable **l_le_wb_pop** in its original scale.
- **lgdp_c15**: The variable **l_lgdp_c15** in its original scale.
- **log_unit_per_capita**: The logged number of smallest units normalized by population.
- **unit_per_capita**: The variable **log_unit_per_capita** in its original scale.

Dataset #2: mainData_yearly.csv

This is the dataset that is used for robustness analysis. The unit of observation is country - eletion-type - year. The variables included in this dataset are mostly the same as above. Some differences are:

1. This dataest has variable **year** in replacement of **period_num** or **period_char**.

[1] Psephos: Adam Carr's Electoral Archive

[2] Link: https://population.un.org/wpp/

2. This dataset doesn't have variables **ecav_violence**, **pre_violence**, or **post_violence**.

3. This dataset has the following additional variables providing detailed descriptive information about the most disaggregated administrative level at which countries publish their electoral results.

    - **unit_level**: a numeric variable indicating admin-level. For example, **unit_level** being 0 means country (admin-level 0) level.
    - **name_english**: the name (in English) of the most disaggregated administrative level.
    - **nlevel**: the number of units at this most disaggregated administrative level at which electoral results are published.

Dataset #3: Provenance Dataset.csv

The provenance data describes and documents the collection process and details of our main datasets. The Provenance Dataset.csv includes the following variables:

- **Country**: name of country.
- **Year**: the year election was held.
- **Election type**: Presidential or Legislative.
- **Date**: Date election was held.
- **Link**: Link (URL) to results page.
- **Domain**: The domain where electoral results were found.
- **DocumentSource**: the authority releasing the information
    - EC: Elections Commission
    - EC—M: Elections Commission Microsite (i.e., a website that is attached to the elections commission website but has a unique subdomain)
    - CC: Constitutional Court
    - IM: Interior Ministry
    - IS—G: Information Service or Gazette
    - SA—DP: Statistics Agency or general Data Portal
    - OD: Other Department
- **Language**: The language the document is published in

- **Site Divisions**: this variable captures the number of times the results are divided until the highest level documentation is available. For example, if navigating to a results document requires selecting a province, then a municipality, then a district, that is three divisions.
- **Data Labels**:
  - Full: The results columns/rows are clearly and explicitly labeled (e.g., "Valid Votes").
  - Partial: The results columns are not clearly labeled but are still decipherable.
  - Encoded: The results columns are labeled only via numerical codes which are meaningless without a key (e.g., putting valid votes and invalid votes in the same column, and using a different column with a numerical key to separate them)
  - None: The columns are completely unlabeled, and their meaning needs to be deduced by analyzing the data.
- **Data Legibility**:
  - MachineReadable: The data is explicitly encoded and can be read by a machine.
  - NotMR: Not machine readable.
  - Corr: Data that is not machine-readable and which is difficult to read. This includes low-quality photocopies and poorly handwritten results.

Dataset #4: vdem.csv

This dataset is the processed V-Dem data that matches the country - year observations in our main datasets. We use this dataset to produce the scatter plots in Supplementary Information – SI-Figure I2, I3, and I4. These plots using raw V-Dem measures are used to illustrates the trend of electoral malpractice over time.

- **country_name**: The country names.
- **year**: The year information.
- **v2elirreg**: The response to the question whether there was "evidence of other intentional irregularities by incumbent, opposition parties, and/or vote fraud" (V-Dem Codebook v11.1 p66). It takes values from 0 (Yes, there were systematic and almost nationwide other irregularities) to 4 (None, there was no evidence of intentional other irregularities). The "Other voting irregularities" outcome variable used in the main dataset is taken and transformed from this raw V-Dem variable.

- **v2elwestmon**: The response to the question "were there allegations of significant vote-fraud by any Western monitors" (V–Dem Codebook v11.1 p62). The variables takes 1 if the answer is yes and 0 otherwise. The "Monitors-reported fraud" variable in the main dataset is taken and transformed from this raw V-Dem variable.

- **v2elfrfair**: The response to the question "Taking all aspects of the pre-election period, election day, and the post-election process into account, would you consider this national election to be free and fair" (V–Dem Codebook v11.1 p70). It takes values from 0 (the election were fundamentally flawed) to 4 (Yes, the election could be considered free and fair). The "Election fraud" variable in the main dataset is taken and transformed from this raw V-Dem variable.