The Subnational Electoral Coercion in India (SECI) Data Set, 1985-2015

Codebook

Cécile Richetta[1]*, Imke Harbers[2], Enrike van Wingerden[3]

[1]Political Science and International Relations, University of Geneva, Geneva, Switzerland
[2]Political Science, University of Amsterdam, Amsterdam, Netherlands
[3]London School of Economics and Political Science, London, UK
*Corresponding author. Email: cecile.richetta@unige.ch

When using this data, please always cite:

Richetta, C., Harbers, I., & van Wingerden, E. 2023. The Subnational Electoral Coercion in India (SECI) Data Set, 1985-2015. *Electoral Studies* Volume(Issue): pp.

When appropriate, also cite this codebook: Richetta, C., Harbers, I., & van Wingerden, E. 2023. The Subnational Electoral Coercion in India (SECI) Data Set, 1985-2015. Codebook. https://doi.org/10.7910/DVN/SCGY8A

Contents

Overall presentation 3
Data collection 3
Time period 3
Event Extraction Procedure 4
ProQuest 4
Factiva 4
Lexis-Nexis 4
Checklist 5
Data sets and Variables 5
Data sets information 5
Variable: unit 6
Variable: state_id 6
Variable: state_name 6
Variable: district_name 6
Variable: constituency_ECI_ID 6
Variable: constituency_name 7
Variable: year 7
Variable: polling_start 7
Variable: polling_end 7
Variable: result_date 8
Variable: fraud 8
Variable: boycott 8
Variable: death_toll 8
Variable: event_description.# 8
Variable: art_name.# 9
Variable: source.# 9
Variable: ecotimes.# 9
Variable: source_elec.# 9
Variable: affected_constituencies 10
Variable: total_constituencies 10
Variable: toi_articles 10

Overall presentation

Please refer to the following article for an introduction of the data:

- Richetta, C., & Harbers, I., & van Wingerden, E. 2023. The Subnational Electoral Coercion in India (SECI) Data set, 1985-2015. *Electoral Studies* Volume(Issue): pp.

For an application of the data, please see:

- Harbers, I., Richetta, C., & Van Wingerden, E. 2023. Shaping Electoral Outcomes: Intra- and Anti-systemic Violence in Indian Assembly Elections. *British Journal of Political Science*, 53(2), 424-440.

The codebook presents the data collection and variable operationalization processes.

Data collection

We code violations of the principle of clean elections in the form of incidents that raise questions about the ability of voters to cast their ballot without fear or intimidation, or that suggest that ballots might not be counted fairly. Negative campaigning (verbal attacks on opponents or name-calling) as well as suggestions of corruption or violations of campaign spending rules do not rise to this level and are not included.

Incidents of violence are coded if they are directly related to the election and are substantiated, i.e., either presented as a fact by the *Times of India* (TOI) or substantiated by the Electoral Commission of India (ECI). This includes violent attacks against voters, poll workers, election officials, candidates or police forces deployed to safeguard the election. Violent incidents that occur prior to the election and that are debated during the campaign, but that do not appear to be directly related to polling or the election, are not included. We also code reports about voter intimidation, or latent evidence for problems during the election when election-related articles mention armed groups that have a history of disrupting polling as having issued a poll boycott.

Time period

We collected data for the period starting one month prior to the first poll and one month after the results are announced.

We extracted polling date(s) (start and end) and result date(s) for each election that took place between 1985 and 2015 in each Indian state from the Electoral Commission of India website https://eci.gov.in. For most elections, polling days happened in one or two days; however, when re-polls are ordered, the dates of these repolls are taken as the end date of polling, and the result date(s) of these re-polls are taken as the overall result date(s).

We used three search engines:

1. ProQuest: 1985-1997
2. Factiva: 1997-2009
3. Lexis Nexis: 2010-2015

Event extraction procedure

Below are the instructions for how to extract articles from each of the news aggregation services we followed.

ProQuest

1) Using the advanced search in the "ProQuest Historical Newspapers: The Times of India" database, use the Command Line option to enter the state/union territory name.
2) Add row to enter additional search terms. In first row, search for: "state name" AND "election". In second row, select the 'AND' option from the drop-down menu to search for: AND '("repoll" OR "re-poll" OR "irregularit*" OR "violen*" OR "unrest" OR "attack" OR "intimidate*" OR "booth captur*" OR "boycott*"). "riot" is omitted as it is incorrectly recognized by ProQuest, yielding articles with the search term "not" instead.
3) Select "specific date range" under "publication date" to select the appropriate date range (one month before first polling date and one month after election results are reported) based on election dates in the data base.
4) Select all results and download them to a local folder "ProQuest". Create a new folder with the name of the state or union territory. Save the document generated by ProQuest with the name of the year of the election (e.g., 2015).
5) Record the number of articles extracted.

Factiva

1) Select the Search Form option in Factiva, enter the state/union territory name in the "this exact phrase" box.
2) Add search term in the "all of these words" box, enter: "election".
3) Add additional terms in the "at least one of these words" box, enter: "repoll, re-poll, irregularit?, violen?, unrest, attack, intimidate?, booth captur?, boycott?"
4) Change the date mode to "Enter date range…" and select the appropriate date range (one month before first polling date and one month after election results are reported) based on election dates in the data base (where available).
5) Select the source by opening the source menu. Enter "Times of India" and select the first option "Times of India – All sources".
6) Select all results and download them to a local folder "Factiva". If necessary, create a new folder with the name of the state or union territory. Save the document generated by Factiva with the name of the year of the election (e.g., 2015).
7) Record the number of articles extracted.

Lexis-Nexis

1) Using the power search interface in Lexis-Nexis, use the HLEAD option to search for the state/union territory name in the headline and lead paragraphs.
2) Use Boolean options to select additional terms. After selecting the country, search for "HLEAD(state/union territory name) AND election AND repoll OR re-poll OR irregularit! OR violen! OR unrest OR attack OR intimidat! OR booth captur! OR boycot!" as additional terms anywhere in the document.

3) Go to "Select Sources" window of the power search and select “Find Sources”. In the “Find Sources” field, select the "News" checkbox and locate “Times of India”.
4) Select appropriate date ranges (one month before first polling date and one month after election results are reported) based on election dates in the data base (where available).
    - It is easy to make a mistake when entering the dates, so be sure to double-check once you have downloaded the articles in .rtf (the search parameters are included on the first page).
5) Select all results and download them to a local folder “Lexis Nexis”. If necessary, create a new folder with the name of the state or union territory. Save the document generated by Lexis Nexis with the name of the year of the election (e.g., 2015).
6) Record the number of articles extracted.

Checklist

- ProQuest and Factiva highlight search terms in yellow in the document, and Lexis Nexis in red. However, ProQuest is inconsistent in highlighting the terms and for some elections none of the terms were highlighted, or only “election” and the name of the state.
- *Verify state*: Is this article really about subnational elections in the state you are coding? Sometimes, incidents in neighboring states are mentioned, since they might influence elections, or multiple states hold elections around the same time and all are discussed in the same article.
- *Verify election*: Is this article about subnational legislative elections (Vidhan Sabha) rather than general or national elections (Lok Sabha), or local (Panchayat) elections?
    - Note that in some instances polling may be held concurrently, so it is impossible to know to which election the article refers. In this case, code the incident.

Data sets and Variables

Data sets information

We identify assembly constituencies affected by electoral coercion and provide information describing acts of coercion. The main data set is in the file named “**seci-main-data-set.csv**”. It contains all information available at the assembly constituency level. Each line in the main data set is a unique assembly constituency – election observation (i.e., each assembly constituency is mentioned once per election). Only assembly constituencies affected by coercion in the election appear in our data. If one article reports poll violence in multiple constituencies, and mentions these by name, then we code violence as having occurred in all these constituencies. Each constituency then is a separate row in the dataset.

Some reports of electoral coercion do not include the constituency: in that case, we move to the next best unit of analysis: district level. Reports of coercive acts are stored in a separate file named “**seci-district-data-set.csv**”. Finally, if neither constituency nor district names are mentioned, we use the state-level and store reports in another distinct file named “**seci-state-data-set.csv**”. The additional data set, “**seci-election-information.csv**” includes information at the election level, such as polling dates and the number of articles extracted. SECI consists of four distinct files:

1. **seci-main-data-set.csv**: main data set at the assembly constituency level.

2. **seci-district-data-set.csv**: complementary data set at the district level.
3. **seci-state-data-set.csv**: complementary data set at the state level.
4. **seci-election-information.csv**: file including information about the elections.

All data sets are provided in ".csv" file format, with comma "," delimiter. The data sets can be merged using the variables "state_name" and "year".

Variable: unit

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv

In the main data set: Unique identifier for each observation, which is a concatenation of the variables (1) state_id (2) constituency_ECI_ID, and (3) year; separated by the operator "-". Format: "00-000-0000".

In the district data set: Unique identifier for each observation, which is a concatenation of the variables (1) state_id, (2) district_name, and (3) year; separated by the operator "-". Format: "00-District-0000".

In the state data set: Unique identifier for each observation, which is a concatenation of the variables (1) state_id, (2) year; separated by the operator "-". Format: "00-0000".

Character format.

Variable: state_id

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv

Unique identifier of state names, comparable across SECI data sets containing incidents.

Character format "00".

Variable: state_name

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv; seci-election-information.csv

Name of the state, comparable across all SECI data sets.

Character format.

Variable: district_name

Data sets: seci-district-data-set.csv;

Name of the district, according to the *Times of India* news report.

Character format.

Variable: constituency_ECI_ID

Data sets: seci-main-data-set.csv;

Official assembly constituency number, according to the Electoral Report of the election published by the Electoral Commission of India. Articles in the *Times of India* usually report the names of the constituencies affected by electoral coercion: we proceeded to manually identify the official number of the assembly constituency in the Electoral Reports, using their names as reported by the *Times of India*.

Numeric format.

Variable: constituency_name

Data sets: seci-main-data-set.csv;

The name of the assembly constituency. If the name is the same in the *Times of India* and the Electoral Reports of the Electoral Commission of India, this name is reported. If the spelling was different between the *Times of India* and the Electoral Reports, the name in the Electoral Reports is reported instead.

Character format.

Variable: year

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv; seci-election-information.csv

Year when the election was held. If the result dates are in a different year than the polling dates, the year when the polling started is taken as the election year.

Comparable across all SECI data sets.

Numeric format.

Variable: polling_start

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv; seci-election-information.csv

Date when the polling started, according to the Electoral Commission of India. Format: day-month-year (e.g., "June 16th, 1991" is "16-06-1991").

Character format.

Variable: polling_end

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv; seci-election-information.csv

Date when the polling ended, according to the Electoral Commission of India. If a re-poll was ordered, the date of the repoll is taken as the end date of the poll. Format: day-month-year (e.g., "June 16th, 1991" is "16-06-1991").

Character format.

Variable: result_date

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv; seci-election-information.csv

Date when the official results are announced, according to the Electoral Commission of India. If there was a re-poll, the date when the results of the repoll are announced is taken as the result date. Format: day-month-year (e.g., "June 16th, 1991" is "16-06-1991").

Character format.

Variable: fraud

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv;

Binary [0;1]. 0 = there was no coercive fraud reported in the assembly constituency; 1 = there was coercive fraud reported in the assembly constituency. The fraud is identified in (one of) the event description variables.

Numeric format.

Variable: boycott

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv;

Binary [0;1]. 0 = there was no boycott call by a non-state armed group reported in the assembly constituency; 1 = there was a boycott call by a non-state armed group reported in the assembly constituency. The boycott call is identified in (one of) the event description variables.

Numeric format.

Variable: death_toll

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv;

Overall number of fatalities linked to electoral coercion during the time frame (1 month prior to the polling days and 1 month after the result dates). When the article mentioned several deaths in multiple constituencies (e.g. "8 deaths were reported from the following constituencies: […]"), then the total number of deaths was divided by the number of constituencies reported, and the final number was entered for each constituency mentioned.

Numeric format.

Variable: event_description.#

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv;

String variable briefly describing the event as reported in the article, as specific as possible. When several reports are made within the same article, they are separated by the operator "[…and…]".

is a number in the data set, from 1 to 5. If reports are made in different articles, they are stored in separate columns that are numbered.

Character format.

Variable: art_name.#

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv;

Titles of the article that mentions the event. Sometimes, the same information will appear in different articles, even word-by-word. This occurs for instance if the information is reported in multiple sections of the newspaper (e.g., regional and national content). Only the first description of the event and first article name are reported.

is a number in the data set, from 1 to 5. If reports are made in different articles, they are stored in separate columns that are numbered.

Character format.

Variable: source.#

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv;

Binary [0;1]. Where possible, we complemented incomplete information about assembly constituencies where incidents happened with alternative national news sources such as *The Hindu* and *India Today*. If the variable "source_elec.#" =1, the name of the assembly constituency was extracted from *The Hindu* or by *India Today*.

is a number in the data set, from 1 to 5. If reports are made in different articles, they are stored in separate columns that are numbered.

Numeric format.

Variable: ecotimes.#

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv;

Binary [0;1]. Factiva gave results from both the *Times of India* and the *Economic Times*. Some events were only reported in the *Economic Times*. We decided to keep these articles, however an additional variable was added to control for a possible bias in these elections, as two sources were used. If the article in "article name" is from the *Times of India*, a 0 was entered. If the article was from the *Economic Times*, a 1 was entered.

is a number in the data set, from 1 to 5. If reports are made in different articles, they are stored in separate columns that are numbered.

Numeric format.

Variable: source_elec.#

Data sets: seci-main-data-set.csv; seci-district-data-set.csv; seci-state-data-set.csv;

Binary [0;1]. For the period covered by ProQuest, some of these articles were found in the results of the extraction process for elections taking place in other states at the same period. If important information was found in these articles, we decided to keep them and add a variable to identify electoral coercion that was identified from the original search results (= 0) or from another election search results (= 1).

is a number in the data set, from 1 to 5. If reports are made in different articles, they are stored in separate columns that are numbered.

Numeric format.

Variable: affected_constituencies

Data sets: seci-election-information.csv;

Number of unique assembly constituencies affected in the state during the election.

Numeric format.

Variable: total_constituencies

Data sets: seci-election-information.csv;

Number of unique assembly constituencies in the state.

Numeric format.

Variable: toi_articles

Data sets: seci-election-information.csv;

Number of articles extracted with our key terms for the election.

Numeric format.