Codebook for "Dataset on Individuals Serving in State Legislatures in the United States, 1900–2016"

Ivan Fomichev and Miriam Golden

2025-01-19

Contents

1 Acknowledgments **2**

2 Funding **3**

3 Overview **3**

3.1 Aims of the project 3

3.2 Variables collected 3

4 General instructions **4**

4.1 Selection criteria 4

4.2 States and legislative periods included 5

4.3 Sources 5

5 Data collection procedures **9**

5.1 General information 9

Codebook References **17**

Dataset description

Assembles individual-level data on all legislators serving in state legislatures in the 50 U.S. states over time. Characteristics collected include first and last name, legislative period, whether reelected into the subsequent legislature and, for 9 states, whether the individual sought reelection into the next legislature. Data is for the lower house or the unicameral assembly. Years generally span 1900 to the late-1940s but in some cases go as far as 2016.

Principal Investigators

Miriam Golden, Peter Mair Chair of Comparative Politics, European University Institute and Research Professor, University of California at Los Angeles

Eugenia Nazrullaeva, Post-doctoral Reseacher in Comparative Politics, University of Konstanz

Contact

Miriam Golden: email `golden@ucla.edu`; homepage https://www.miriamgolden.com/

Codebook Authors

Ivan Fomichev, European University Institute

Miriam Golden, European University Institute and University of California at Los Angeles

How to Cite this Dataset

Miriam A. Golden and Eugenia Nazrullaeva, 2025. "Dataset on Individuals Serving in State Legislatures in the United States, 1900–2016." DOI: https://doi.org/10.7910/DVN/UCR2AU, Harvard Dataverse.

1 Acknowledgments

For assistance with data collection, we thank Sohaila Abdelaal, Jacques Courbe, Brian Hamel, Diala Husseini, Siyu Liang, Hailee Nguyen, and Ali Abdul Rehman Unar.

2 Funding

We gratefully acknowledge funding from the Academic Senate of the University of California, the Research Council of the European University, and the Department of Economics at the University of Warwick.

3 Overview

3.1 Aims of the project

The aims of this project are to assemble data on individual legislators serving in U.S. state legislatures over the entire 20^{th} centery and into the 21^{st} in order to calculate rates of reelection and, to the extent possible, rerunning rates. Existing datasets of elected legislators that we consulted and supplemented are Ansolabehere, Ban, and Snyder (2017) and Klarner (2018). The first mainly covers the period from 1920 to 1968 and the second covers 1967 to 2016. To calculate complete reelection rates by legislature-election, we collected information on all legislators who were missing from both these other datasets for any legislature between 1900 and 2016.

The posted dataset includes only the data collected by Golden and Nazrullaeva for the years not available in Ansolabehere, Ban, and Snyder (2017) or Klarner (2018). Researchers seeking data on other years should consult either Ansolabehere, Ban, and Snyder (2017) or Klarner (2018).

All rerunning data was collected specifically for the current project and is included in the posted dataset.

3.2 Variables collected

Variables collected are: legislators' first and last names, the legislative period (the year the legislature was convened), and whether the individual was an incumbent in the election at time t. In some cases, we were able to retrieve party affiliation and the constituency of legislators, although these data were not systematically collected.

For nine states, we also collected data on whether incumbent legislators chose to run again for the subsequent election at time $t + 1$. Additionally, the dataset contains quality control measures used to assess the accuracy of rerunning coding. Table 1 lists all the variables included in the dataset.

Table 1: Variables in the Dataset

Variable	Description
state	U.S. state abbreviation.
year	Year of the election.
id	ID of the representative.

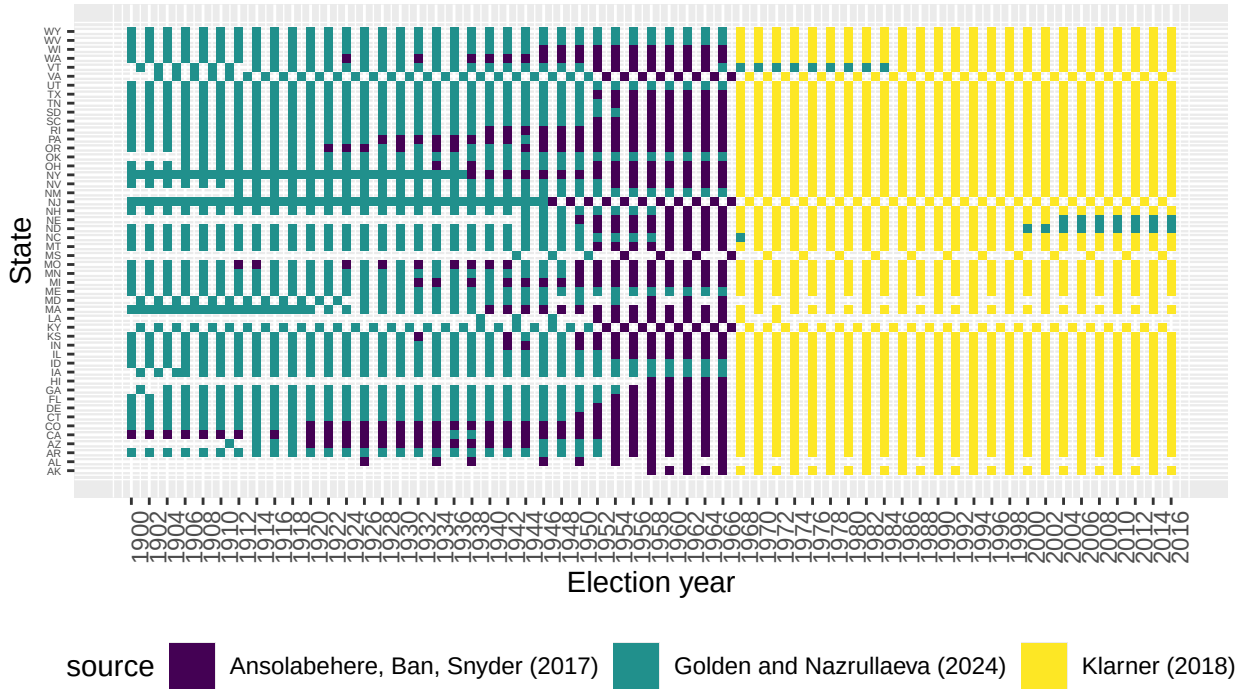
Table 1: Variables in the Dataset *(continued)*

Variable	Description
name_full	Full name of the representative.
last	Last name of the representative.
first	First name of the representative.
middle	Middle name or initial of the representative (if available).
party	Political party affiliation of the representative (if available).
county	The county represented by the legislator (if available).
inc	Binary indicator for incumbency status at the moment of the election at time t. 1: incumbent, 0: non-incumbent.
runs_next_election	Binary indicator for whether the representative reran for election at time t+1. 1: reran, 0: did not rerun.
ocr_match_same_year	Measure to check if the representative's name could be found in the register of votes for the same election year t. Proxy for OCR quality and the accuracy of rerunning coding.
rerun_mismatch	Binary indicator for representatives who won in t+1 but were mistakenly classified as not rerunning.

4 General instructions

4.1 Selection criteria

Aiming to fill the gaps in Ansolabehere, Ban, and Snyder (2017) and Klarner (2018), we included all the legislatures in the period from 1900 to 2016 that were not included in either. The graph below shows which state-years are included in the current dataset and which by Ansolabehere, Ban, and Snyder (2017) or Klarner (2018).



4.2 States and legislative periods included

4.3 Sources

The sources for the data that we collected include legislative journals, legislative manuals, reports compiled by legislative research commissions, official legislature websites, and official state registers. Where information could not be acquired online, we collected data through direct communication with states' legislative assembly offices, archives, and libraries.

Table 2 lists the sources of our data by state and indicates whether the data were manually processed or automatically extracted into a tabular format suitable for analysis.

Table 2: Years Covered, Data Sources, and Data Extraction Techniques for All States

state	years	source_name	source_type	method
Arizona	1946-1952	Arizona Memory Project	Website	Manual
Arizona	1911-1918	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Arkansas	1901-1945	Historical Report Of The Secretary Of State, 2018	Scan	Automatic (R/Python)
Arkansas	1947-1953	Historical Report Of The Secretary Of State, 2018	Scan	Manual

California	1914, 1918, 1936, 1938	Journal Of The Assembly	Scan	Automatic (R/Python)
Colorado	1913-1917	Colorado Legilsators Past And Present	Website	Manual
Colorado	1901-1911, 1919	House Journal Of The General Assembly	Scan	Automatic (R/Python)
Connecticut	1898-1948	Connecticut General Assembly Members	Website	Automatic (R/Python)
Delaware	1901-1945	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Delaware	1947-1951	Journal Of The House Of Representatives	Scan	Manual
Florida	1898-1952	Membership Of The Florida House Of Representatives By County	Scan	Automatic (R/Python)
Georgia	1940-1954	Georgia's Official Register	Scan	Manual
Georgia	1899-1939	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Idaho	1890-1966	Idaho Blue Book	Scan	Automatic (R/Python)
Illinois	1903-1913, 1945-1953	Blue Book Of Illinois	Scan	Manual
Illinois	1901-1903, 1913-1945	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Indiana	1901-1941	House Journal	Scan	Automatic (R/Python)
Indiana	1947-1949	House Journal	Scan	Manual
Iowa	1837-1966	Historical Tables Of The Iowa Legislature	Website	Automatic (R/Python)
Kansas	1901-1911, 1935-1943	House Journal	Scan	Automatic (R/Python)
Kansas	1913-1931, 1943-1949	House Journal	Scan	Manual
Kentucky	1900-1951	Kentucky General Assembly Membership 1900-2005	Scan	Manual

Louisiana	1939, 1943, 1947	Membership In The Louisiana House Of Representatives 1812 - 2020	Scan	Manual
Maine	1819-1966	Maine State Law & Legislative Reference Library	Xlsx	Xlsx
Maryland	1901-1945	Archives Of Maryland, Historical List House Of Delegates, 1790-1990	Website	Automatic (R/Python)
Maryland	1947-1955	Journal Of The House Of Representatives	Website	Automatic (R/Python)
Massachusetts	1900-1951	Wikipedia, Xxxx Massachusetts Legislature	Website	Automatic (R/Python)
Michigan	1900-1930, 1936, 1940	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Minnesota	1934-1948	General Election Returns For Minnesota	Scan	Manual
Minnesota	1900-1932	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Mississippi	1943, 1947, 1951	Hand Book : Biographical Data Of Members Of Senate And House, Personnel Of Standing Committees	Scan	Manual
Missouri	1899-1935	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Missouri	1945-1949	Journal Of The House Of Representatives	Scan	Manual
Montana	1901-1943	House Journal Of The Legilsative Assembly	Scan	Automatic (R/Python)
Montana	1945-1951	House Journal Of The Legilsative Assembly	Scan	Manual
Nebraska	1944-1948, 2004-2016	Nebraska Blue Book	Scan	Manual
Nevada	1900-1952	Journal Of The Assembly	Scan	Manual
New Hampshire	1900-1927	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
New Hampshire	1929-1960	State Archives Office	Xlsx	Xlsx

New Jersey	1898-1946	Minutes Of Votes And Proceedings Of The General Assembly Of The State Of New Jersey	Scan	Automatic (R/Python)
New Mexico	1922-1966	New Mexico State Legislature Legislative Council Service	Native Pdf	Automatic (R/Python)
New York	1901-1938	Journal Of The Assembly	Scan	Automatic (R/Python)
North Carolina	1969	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
North Carolina	1900-1942	North Carolina Government, 1585-1974: A Narrative And Statistical History. Edited By John L. Cheney, Jr	Scan	Automatic (R/Python)
North Dakota	1900-1917, 1921-1937, 1941-1943, 1919, 1939	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
North Dakota	1945-1951	Journal Of The House Of Representatives	Scan	Manual
Ohio	1900-1910, 1925-1933, 1937-1945	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Ohio	1947-1951, 1912-1922	Journal Of The House Of Representatives	Scan	Manual
Oklahoma	1906-1966	Historic Members	Website	Automatic (R/Python)
Oregon	1900-1920, 1928, 1932, 1938-1942	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Pensylvannia	1901-1927, 1945	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Rhode Island	1901-1929	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Rhode Island	1930-1938	Journal Of The House Of Representatives	Scan	Manual
South Carolina	1901-1943	Journal Of The House Of Representatives	Scan	Automatic (R/Python)

Tennessee	1945	House Archives - Tn General Assembly	Scan	Manual
Tennessee	1898-1944, 1946-1952	House Archives - Tn General Assembly	Website	Automatic (R/Python)
Texas	1900-1950	Texas Legislators: Past & Present	Website	Manual
Utah	1895-1966	Utah State Legislature, Legislators By Year	Website	Manual
Vermont	1900-1943, 1967-1985	Journal Of The House Of Representatives	Scan	Automatic (R/Python)
Vermont	1945-1951	Journal Of The House Of Representatives	Scan	Manual
Virginia	1900-1935	A History Of The Virginia House Of Delegates	Website	Automatic (R/Python)
Virginia	1937-1951	A History Of The Virginia House Of Delegates	Website	Manual
Washington	1901-1937	House Journals Of The State Of Washington	Scan	Automatic (R/Python)
West Virginia	1900-1966	West Virginia Archives And History	Website	Automatic (R/Python)
Wisconsin	1916-1944	Journal Proceedings Of The Wisconsin Legislature	Scan	Automatic (R/Python)
Wisconsin	1900-1914	Wisconsin Blue Book	Scan	Manual
Wyoming	1890-1966	Wyoming Legislator Database	Website	Manual

5 Data collection procedures

5.1 General information

Data were collected between 2020 and 2023.

5.1.1 Digitizing archival records in R

In total, the data for 1309 legislative periods in 47 states were extracted from scanned (.pdf) journals of the lower house or other historical documents. Where the scan quality allowed, we digitized such documents using primarily the programming environment R and, to a lesser extent, Python. In such cases, the procedure was the following:

1) First, we identified and selected pages in the PDF documents that contained the lists of assembly members. This list was either in the form of a dedicated table or as a roll call of the representatives in the first legislative session. In cases of roll call lists, we sometimes missed legislators who were absent in the first session, but the prevalence of such cases was under 5%.

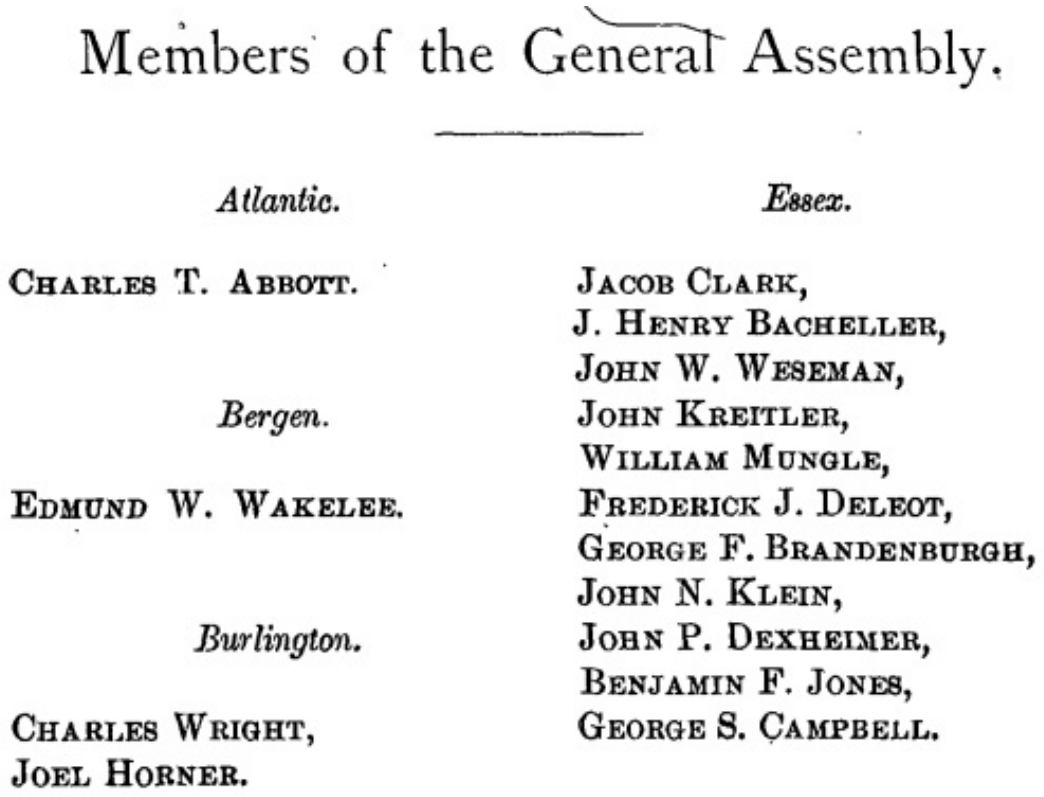
2) Second, we extracted text line by line from the specified pages and combined them in a data frame. If the PDF document did not contain a recognized text layer, or if the latter was incomplete or contained too many inaccuracies, the Tesseract OCR engine was used to digitize text from the images.

3) Third, we extracted legislators' names from rows containing raw text. The main task here was to identify which part of a string contained a name and what repeated pattern could be used to separate it from the rest of the string. Where possible, we sought to extract other legislator characteristics, such as party affiliation and/or county where elected.

4) Fourth, we cleaned and standardized names to correct misspellings and formatting inconsistencies. That is, we checked for consistency in names' format, ensuring first-middle-last name order is the same across states for all legislators. As we were dealing with scanned versions of documents without the native text layer, we also encountered occasional misspellings. To deal with them, we:

    a) Checked for atypical characters in legislators' names and manually corrected those. In particular, we checked for the presence of uncommon non-letter characters, two punctuation symbols in a row, two upper-case letters in a row, two letters that rarely go together (such combinations as kk, xx, yy, hh), a white space not followed by an upper-case letter or two lower-case letters separated by a white space etc. This method, however, did not allow to spot less obvious misspellings, such as "Pattereon" instead "Patterson", for example.
    b) To address those types of errors, we then compared legislators' last names to the list of the most common 151,000 last names in the US. If a legislator's last name was not present in the list, we first searched for a legislator elected in a different year with a surname that differed from that suspicious last name by no more than one character and was present in the list of the most common last names. If such a name was found, we replaced our original suspicious name with it. If there was no such name, we manually checked the spelling, and, if necessary, corrected it. For example, suppose we have a legislator with the surname "Pattereon" elected in 1901. This surname is misspelled and therefore not present in the list of the most common last names. We then search for a similar last name among legislators elected at a year $\neq 1901$ and, finding "Patterson" among them, replace the original last name with it.

As an example, below is an extract from the legislative journal of the state of New Jersey, a typical pdf document we worked with to extract the names of legislators.



Members of the General Assembly.

Atlantic.

CHARLES T. ABBOTT.

Bergen.

EDMUND W. WAKELEE.

Burlington.

CHARLES WRIGHT,
JOEL HORNER.

Essex.

JACOB CLARK,
J. HENRY BACHELLER,
JOHN W. WESEMAN,
JOHN KREITLER,
WILLIAM MUNGLE,
FREDERICK J. DELEOT,
GEORGE F. BRANDENBURGH,
JOHN N. KLEIN,
JOHN P. DEXHEIMER,
BENJAMIN F. JONES,
GEORGE S. CAMPBELL.

Table 3 demonstrates how this page is represented as a data frame in R (raw data, the left column). Then, after some data cleaning and manipulation — separating columns by a pattern, splitting rows with more than two names etc. — legislator names appear as in the right column.

As one can see, there is a spelling error in the name "George S. Campbell," whose middle initial was digitized as "8" instead of as "S" (a common issue due to the similarity between the two symbols). This was dealt with at the last stage of the data preparation process when spelling errors in legislators' names were manually corrected state-by-state.

5.1.2 General Issues with Coding Rerunning and Reeleciton

In general, to identify a legislator as reelected, we need to confirm that the individual was elected to the legislature at time t and at $t-1$. In practice, this means we are looking for the same name in two consecutive legislative journals. Although this task is straightforward yet time-consuming for a human, it can be challenging for a computer for two reasons.

First, the same name can be written differently between two journals, or between the list of elected legislators in a journal and the register of voters which contains the names of all candidates who ran in an election. For example, the person can be recorded as "J. J. Smith" in the register of votes but as "James John Smith" in the corresponding legislative journal. To deal with this problem, we match legislators by their first name initials and surnames. This way, J. J. Smith and James John Smith are considered to be the same legislator. At the same time, it creates the possibility that James John Smith from the

raw_line	cleaned_line
Members of the GeneraT Assembly,	Charles T. Abbott
	Jacob Clark
Atlantia. Essex.	J. Henry Bacheller
	John W. Weseman
Charles T. Abbott. Jacob Clark,	John Kreitler
J. Henry Bacheller,	William Mungle
John W. Weseman,	Edmund W. Wakelee
Bergen. John Kreitler,	Frederick J. Deleot
William Mungle,	George F. Brandenburgh
Edmund W. Wakelee. Frederick J. Deleot,	John N. Klein
George F. Brandenburgh	John P. Dexheimer
John N. Klein,	Benjamin F. Jones
Burlingtm. John P. Dexheimer,	Charles Wright
Benjamin F. Jones,	George 8. Campbell
Charles Wright, George 8. Campbell.	Joel Horner
Joel Horner.	NA

Table 3: PDF text read as raw lines and after cleaning

list of representatives will be matched to Jake John Smith. Although such inaccuracies are possible, their distribution should be random and should not bias the dataset.

Second, since we are mostly working with data extracted from pdf files without a native text layer, there are some spelling errors due to the poor quality of the scans. Because of this, James John Smith can be recorded in our data as Iames John Smith. To address this, we employ fuzzy matching, allowing names that differ by one character to be considered the same. This way, Iames John Smith will be matched to James John Smith. However, it also means that two different names might be inaccurately matched and a legislator wrongly coded as, for instance, reelected. We do not consider this an important problem, since these spelling errors should be minimal in number and distributed randomly.

5.1.3 Coding Reelection

We coded a legislator from state S as reelected from the legislature convened at time t if a person with the same name appears in our dataset among state S legislators at time $t-1$ (i. e. elected in the previous election). To identify reelected representatives, we use first name initial and surname, allowing one character to differ between the names at t and $t-1$ for a given legislator to be coded as reelected.

5.1.4 Coding Rerunning

We identify as rerunners at time t those representatives whose names were found in a list of candidates at time $t+1$. Again, we used initials and last names, allowing one character to be different. This time, however, instead of digitizing registers of votes (or other documents with the candidates' names) into a tidy dataset and then comparing it with the dataset of representatives, we worked directly with raw texts. The reason we do so is because official registers of votes are usually significantly more difficult to digitize and put in the tidy format: primary this is because there are several times more inputs to process in comparison to the list of elected representatives, and also because the layout of a document can be more complex, e.g. with both vertically and horizontally positioned text on the same page.

In general terms, the procedure consisted in checking if a given representative's name from a legislature at a time t could be found in the OCR text layer of a scan of an official register of votes (or other official document of a similar type) that contained the names of candidates who run in the election at time $t+1$. In other words, we took "James John Smith" from the list of representatives elected in 1900 and checked if a person with the same name appeared in the list of candidates in the 1902 election, repeating this procedure for all legislators from the state in a given year.

More specifically, the procedure was the following:

1) Read the register of votes as raw text, remove all non-letter characters, indentations and double white spaces.

2) For each representative from our dataset, take their initials and last name, and search for it in the text. Here we used regular expressions of the form **(J[a-z]{0,10} J[a-z]{0,10} Smith){e<=1}** to account for the fact that first and middle names can be written either as initials or in full format. In other words, J. J. Smith could be matched to James John Smith, James J. Smith, J. J. Smith, or J. J. Smlth.

This approach allows us to calculate rerunning rates relatively fast and without any need to digitize and tabulate registers of votes, but at the cost of not being able to work with low quality OCR text layers containing too many spelling errors. In such cases, the sensitivity of our results would be too low and the proportion of false negatives (representatives who rerun but are not identified as such) too high.

To distinguish between the state-years for which our approach is sensitive enough and those for which it is not, we use two measures. First, we calculate how many representatives from our list elected at time t we were able to find in the register of votes for the **same year**, which serves a proxy for the quality of the scanned data. For example, we took "James John Smith" from the list of representatives elected in 1900 and checked if the algorithm was able to find this name in the list of all candidates in the *1900* election, repeating this procedure for all state legislators in a given year. In an ideal case, for each state-year the number of representatives identified this way should be equal to the number of legislators we have on our list, since all elected representatives should be on the list of candidates. Whenever the number is smaller, it indicates underperformance of the procedure caused by low quality of the text layer. Figure 1 reports the results.

Second, we used reelection, which was coded using cleaner data and a more robust procedure, as an additional benchmark for the accuracy of the rerunning variable. Whoever is coded as reelected in the next election should, by definition, run in that election. Therefore, if a candidate is marked as reelected but not as running in the election, it indicates a likely inaccuracy in the latter coding. Figure 2 reports the results. We do not filter by these variables, but preserve the data as is. We suggest reasonable critical values for these two variables would be 0.8 and 0.05 respectively.

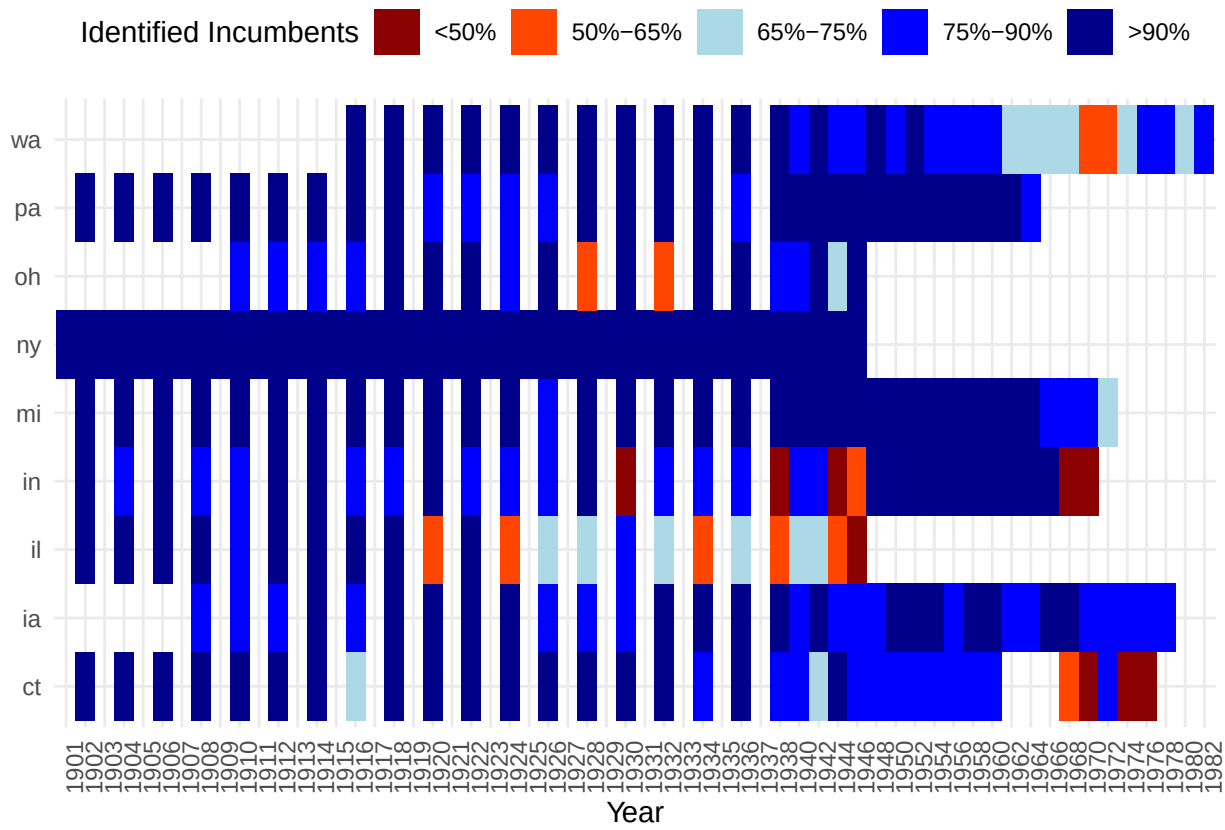


Figure 1: Percentage of Identified Legislators at Time t

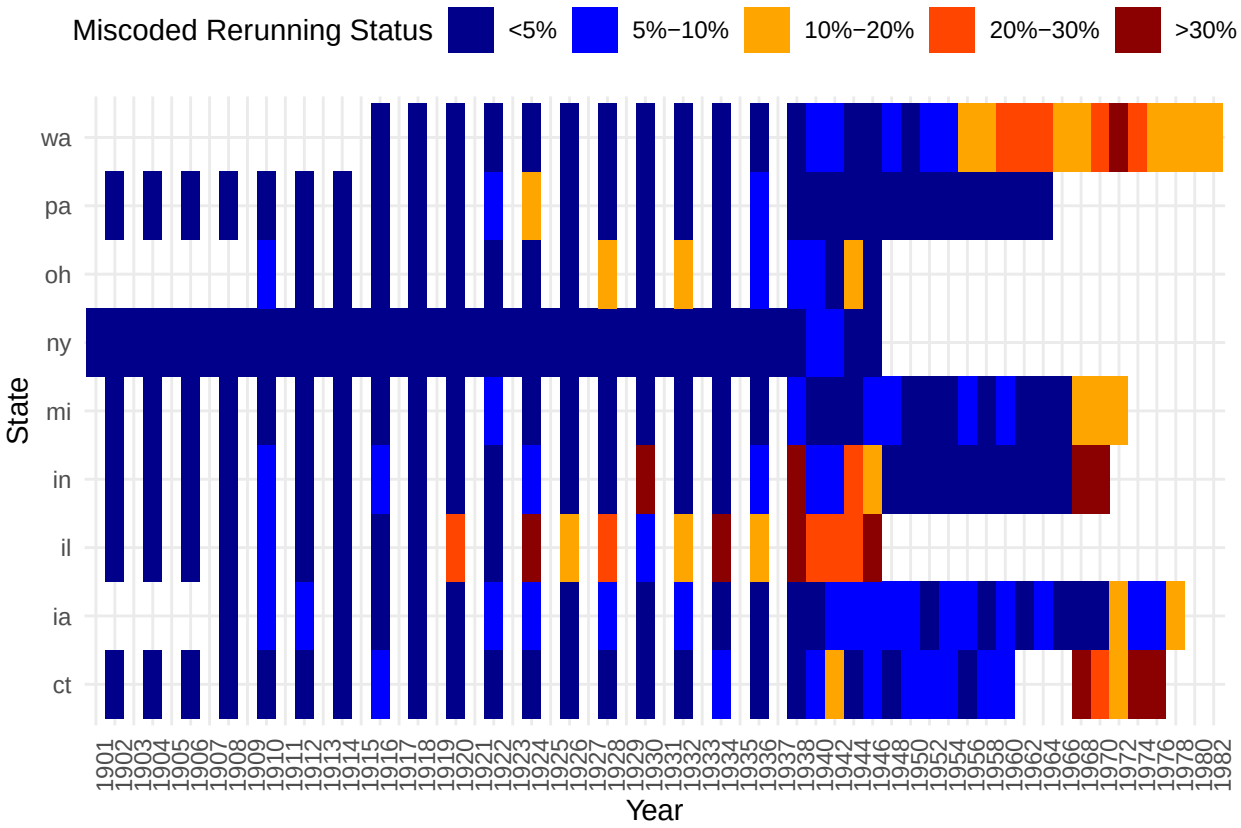


Figure 2: Miscoded Rerunners

Codebook References

Ansolabehere, Stephen, Pamela Ban, and James M. Snyder Jr. 2017. "State Legislative Historical Elections." 2017. https://doi.org/10.7910/DVN/LEMNXZ, Harvard Dataverse, V1, UNF:6:8UQYfDIsmII/tgD+Hrv/8Q== [fileUNF].

Klarner, Carl. 2018. "State Legislative Election Returns, 1967–2016." 2018. https://doi.org/10.7910/DVN/
Harvard Dataverse, V3, UNF:6:pV4h1CP/B8pHthjjQThTTw== [fileUNF].