Leveraging Historical Voter Files as Accurate Measures of Who Votes: Analyzing and Disseminating Voter File Data to Enhance Understanding of Elections

Technical Report
The Evolving Election Administration Landscape Project MIT Election Data and Science Lab

Seo-young Silvia Kim*‡, Bernard L. Fraga†, and Daron Shaw*

‡American University/Sogang University
†Emory University *University
of Texas at Austin

Last Updated Feb 15, 2024

*Corresponding author email: sskim.research@gmail.com. We thank Christopher Roebuck, Jafer Hasnain, Kristen Munnelly, Cornelia Lawrence, and Derek Wakefield for their research assistance.

Contents

1 Introduction

Voter registration databases or voter files are increasingly used for election campaigns, research, and public distribution of election-related statistics. For research, they provide better estimates of who voted, free from non-response and social desirability biases that often inflate turnout estimates reported in surveys (Burden 2000; Ansolabehere and Hersh 2012; Berent et al. 2016; Fraga 2018; Jackman and Spahn 2019). For campaigns and civic organizations, they provide a basis for field operations and experiments (Green and Gerber 2005; Enos et al. 2014). Most importantly, for election administrators, they establish who is eligible to participate in an election (Ansolabehere and Persily 2010; Hersh 2015), record who voted in past elections, and contribute to authoritative, certifiable election reporting.

With changes to election laws due to the COVID-19 pandemic, however, individual-level records of who registers, who votes, and how individuals vote (absentee/mail-in, in-person early, or in-person Election Day) are under increased scrutiny. Unfortunately, voter files are often misunderstood by the public, and this misunderstanding can lead to lowered trust in elections. For example, claims of voter fraud have stemmed from misunderstandings about what voter files are and how they change over time (Grimmer et al. 2023).

Of course, voter file data and analyses do have shortcomings. While voter files offer many improvements over survey data for researchers, the fact that voter registration databases are, primarily, a basic resource for administrators means that one must be careful when assessing them (Igielnik et al. 2018). For instance, using voter file data collected just after the 2004 election, McDonald (2007) found that the average jurisdiction showed 1.8% *fewer* votes from the file than ballots officially recorded in the election. For an inexperienced researcher (or an interested partisan), this discrepancy could serve to undermine one's confidence in the reliability of voter file data. Thus, verifying, clarifying, and improving these records is an important task for social scientists. More to the point, in an election environment increasingly plagued with accusations of fraud and suppression, we need a clear and reliable record of the vote to avoid the further degradation of faith in American elections.

In this report, we seek to build on our experience with voter file data to identify a set of "best practices" in utilizing historical voter files to understand voter behavior and turnout. While several companies (L2, Aristotle, Catalist, etc.) currently compile and maintain a national database of registered voters, to our knowledge no one is systematically archiving these data to create a historical record of registration and voting. From a commercial perspective, this makes *some* sense: for anyone interested in campaign or survey work, the updated list is the only one that matters. From an academic research perspective, as well as for anyone interested in learning about turnout and voting over time, compiling and analyzing data from historical files of registered voters is an untapped potential gold mine. In approaching this task, we use L2's commercial, national archive of raw files from the states, which opens new and

exciting opportunities to expand the breadth and depth of our understanding. We first construct extensive metadata on what information these historical state files contain. This is no small feat, given their variance in format and content. Based on this, as well as our previous experiences, we then identify and discuss "best practices" for generating and acquiring voter file data. Here we consider the perspectives of both "suppliers" (election officials and election administrators) and "consumers" (mainly academics, but also political parties and consultants, civic organizations, and news media).

Beyond compiling and "cleaning" data, we analyze two key research questions. The first question asks whether these raw historical voter files are accurate records of who registers and votes, or whether more complex modeling is necessary. We conclude that raw historical voter files are *fairly* accurate repositories of individual-level voter registration and turnout data. The second question is more specific: how often do voters change *the way* that they vote? To be clear, we are not talking about shifts in their partisan vote choice: rather, we are referring to whether they shift from voting in-person on Election Day to some form of convenience voting, such as absentee/mail or in-person early voting (or vice-versa). On this count, we find that only a small share of voters changes the way they vote from election-to-election. Put a slightly different way, the changes we observed in the relative use of different voting methods are driven by different voters coming into the electorate, rather than existing voters changing their preferred method of voting. This analysis is important to anyone interested in how voting methods might influence voter participation and satisfaction, but is also an excellent example of how cleaned and reliable historical voter files can facilitate relevant professional research.

2 Archived State Voter Files: Availability, Characteristics, and Best Practices

2.1 Data Acquisition from L2 Inc.

In early January of 2023, we used funds from a MEDSL grant to obtain state-level historical voter registration and vote history lists from L2.[1] These lists cover elections from 2004-2023, although the vast majority come from 2016-2023. These lists are raw and unprocessed: they are basically the files obtained from the secretaries of states and have not been cleaned, standardized, or augmented by L2 for public or commercial use. As such, they differ considerably from the statewide and national voter files typically used by academics. The full data set includes at least one registered voter list for each state for every year from 2018-2023.

[1] L2 is a well-established commercial vendor of voter registration lists. Although we have access to these data and are creating our own version of "processed" historical files, L2 retains the underlying rights of ownership. This means that we cannot disseminate or provide access to our data without L2's consent. It is our hope that L2 will be a willing partner in allowing other researchers to validate our research.

2.2 Coverage and Content of Raw Voter Files

2.2.1 Temporal Coverage of Raw Voter Files.

Table 1 and Figure 1 offer comprehensive information about the range and scope of our data sets. For each state, they show (1) the date of the first voter file, (2) the date of the most recent voter file, and (3) the total number of voter files across that range of dates ("snapshots"). As one can see, there is considerable variance by state. There are two reasons for this. First, the data for some states goes back to almost 2000 (2001 for New Jersey, 2003 for North Carolina, and 2004 for Florida, Oregon, Pennsylvania, and Washington), while for other states the data are unavailable until 2018 (Delaware and Virginia). Second, while almost all states update their lists once every six months, other update much more frequently.

State	Min. Date	Max. Date	Number of Snapshots
Alabama	2009 Feb	2022 Dec	22
Alaska	2012 Jul	2023 Jan	21
Arizona	2016 Oct	2023 Mar	12
Arkansas	2010 Sep	2023 Mar	24
California	2009 Oct	2023 Mar	30
Colorado	2012 Mar	2023 Jan	33
Connecticut	2009 Dec	2023 Mar	23
Delaware	2018 Sep	2023 Jan	9
Florida	2004 Aug	2022 Dec	53
Georgia	2010 Sep	2023 Jan	29
Hawaii	2012 Apr	2023 Feb	20
Idaho	2012 Aug	2022 Oct	21
Illinois	2004 Sep	2022 Sep	33
Indiana	2006 Mar	2022 Sep	39
Iowa	2011 Feb	2023 Jan	25
Kansas	2010 Sep	2022 Oct	26
Kentucky	2011 Jun	2023 Jan	22
Louisiana	2010 Sep	2022 Jun	25
Maine	2011 Feb	2023 Mar	23
Maryland	2010 Jun	2022 Sep	22
Massachusetts	2017 Apr	2022 Aug	13
Michigan	2010 Jul	2022 Sep	28
Minnesota	2012 Apr	2022 Oct	16
Mississippi	2011 Feb	2022 Aug	19
Missouri	2009 Jun	2023 Feb	24
Montana	2012 Aug	2023 Jan	20
Nebraska	2009 Dec	2022 Dec	23
Nevada	2012 Jan	2023 Jan	27
New Hampshire	2011 May	2022 Nov	19
New Jersey	2001 Sep	2023 Jan	37
New Mexico	2010 Jul	2023 Feb	18
New York	2011 Jan	2022 Dec	36
North Carolina	2003 Mar	2023 Jan	42
North Dakota	2009 Jan	2023 Mar	19

Ohio	2006 Jul	2023 Jan	34
Oklahoma	2010 Apr	2023 Feb	28
Oregon	2004 Sep	2022 Aug	30
Pennsylvania	2002 Aug	2023 Jan	36
Rhode Island	2010 Sep	2023 Feb	21
South Carolina	2010 May	2023 Feb	25
South Dakota	2012 Apr	2023 Feb	21
Tennessee	2010 Oct	2023 Jan	25
Texas	2010 Jan	2023 Jan	52
Utah	2012 Jul	2023 Feb	20
Vermont	2012 Aug	2022 Sep	23
Virginia	2018 Jun	2022 Sep	11
Washington	2004 Jul	2023 Jan	48
West Virginia	2012 Apr	2022 Mar	22
Wisconsin	2012 Jan	2023 Jan	21
Wyoming	2012 Oct	2023 Mar	18

Table 1: Temporal Coverage of L2 Data by March 2023

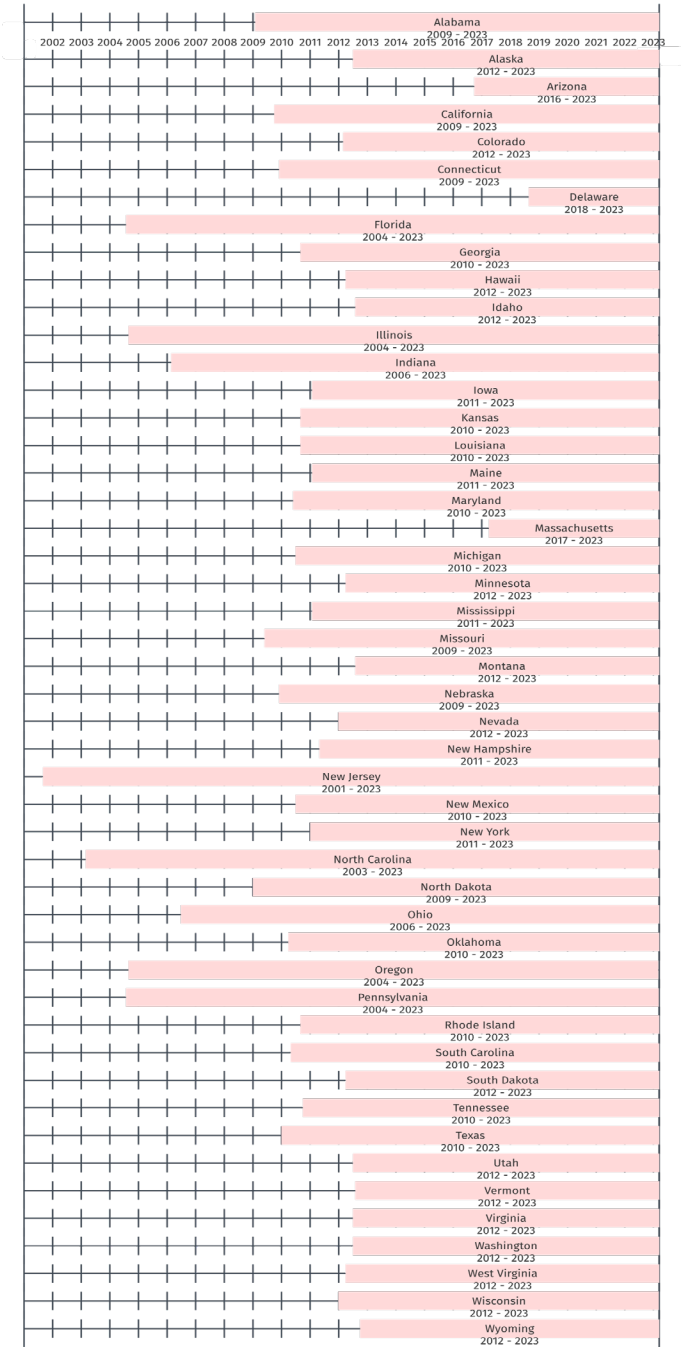


Figure 1: Temporal Coverage of L2 Data by March 2023

2.2.2 Size of State Voter Files and Creating Usable Data Bases

While we were excited to recover these registered voter lists, we faced daunting data storage and data handling issues. Given the need to analyze, assess, and (ultimately) format and standardize hundreds of statewide voter files, we needed a practical approach. Consequently, we chose to keep the statewide datasets separate and to standardize them within each state. This allowed us to perform analyses overtime

within a state, and to then combine the results of these (statewide) analyses to obtain national-level insights. This is akin to the *split-apply-combine* approach adopted by many **R** packages since Wickham (2011). Even so, these statewide voter files often exceeded common computing power: California's raw voter registration list snapshot from April 2023, for instance, stores 280 bytes of data across 53 columns for each of their 25 million registrants on average, occupying 7GB of memory in a tab-delimited file and >18GB of RAM when read into R.[2] Merely understanding what was *stored* in raw voter file snapshots required better-than-average computing capacity. Fortunately, access to personal computers or cloud computing platforms with >64GB of available RAM, along with "lazy" data loading R packages such as data.table, vroom, and readr, allowed us to successfully load (and read) all tabular registered voter data files.

2.2.3 Different Formats of State Voter Files

As even the casual student of U.S. Constitution knows, American voting laws and election administration are left to the states. Because of this, voter file formats vary significantly across the states. We anticipated significant challenges in processing the raw data from historical voter files, but we also assumed that cleaning and standardizing these data would constitute a major contribution for those interested in leveraging voter files to study turnout and participation (Hersh 2015). The reality is that the historical voter lists were less idiosyncratic than we expected. Table 2 shows the format of statewide files, along with some details about information included in the files.

Name	Separate History	Vote Method Available	Status	Orig. Reg. Date	Reg. Date	Party	Gender	Race	Date of Birth
Alabama	No	No	Yes	No	Yes	Yes	Yes	Yes	Date
Alaska	No	Yes	No	Yes	Yes	Yes	Yes	No	No
Arizona	Yes	Yes	No	No	Yes	Yes	No	No	Date
Arkansas	No	Yes	Yes	No	Yes	Yes	No	No	Yes
California	Yes	Yes	No	No	Yes	No	Yes	No	Date
Colorado	Yes	Yes	No	No	Yes	Yes	Yes	No	Year
Connecticut	No	Yes	Yes	No	Yes	Yes	Yes	No	Date
Delaware	Yes	No	No	No	Yes	Yes	Yes	No	Date
District of Columbia	No	Yes	Yes	No	Yes	Yes	No	No	Year
Florida	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Date
Georgia	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Year
Hawaii	No	No	Yes	No	Yes	No	No	No	No
Idaho	No	Yes	No	No	Yes	Yes	Yes	No	No
Illinois	Yes	No	Yes	No	Yes	No	Yes	No	Year
Indiana	Yes	Yes	Yes	No	Yes	No	Yes	No	Date
Iowa	No	Yes	Yes	No	No	Yes	Yes	No	Date
Kansas	No	No	Yes	No	Yes	Yes	Yes	No	Date
Kentucky	No	No	Yes	No	Yes	Yes	Yes	No	Year
Louisiana	Yes	No	Yes	No	Yes	Yes	No	Yes	Year
Maine	Yes	Yes	Yes	No	Yes	No	No	No	Year
Maryland	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Year
Massachusetts	Yes	No	No	No	Yes	Yes	Yes	No	Date
Michigan	Yes	No	Yes	No	Yes	No	Yes	No	Year

[2] This excludes voter history, phone numbers, electoral jurisdiction identifiers, email addresses, and other pieces of voter information that are stored in separate files. Combining all this information, a single "snapshot" from California is an uncompressed 23GB set of distinct tabular data files; 934 bytes (nearly 1MB) are stored per registrant in California's voter registration system.

Minnesota	Yes	Yes	No	No	Yes	No	No	No	Year
Mississippi	Yes	No	Yes	No	Yes	No	No	No	No
Missouri	No	No	Yes	No	Yes	No	No	No	Year
Montana	Yes	No	Yes	No	Yes	No	No	No	Date
Nebraska	No	No	Yes	No	Yes	Yes	Yes	No	Year
Nevada	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Date
New Hampshire	Yes	Yes	No	No	Yes	Yes	Yes	No	Date
New Jersey	Yes	Yes	Yes	No	Yes	Yes	No	No	Date
New Mexico	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Year
New York	Yes	Yes	Yes	No	Yes	No	Yes	No	Date
North Carolina	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Year
North Dakota	--	--	--	--	--	--	--	--	--
Ohio	No	No	Yes	No	Yes	Yes	No	No	Date
Oklahoma	Yes	Yes	No	No	No	Yes	No	No	Date
Oregon	Yes	No	Yes	No	Yes	Yes	No	No	Date
Pennsylvania	No	Yes	Yes	No	No	Yes	Yes	No	Date
Rhode Island	Yes	No	Yes	No	Yes	Yes	Yes	No	Year
South Carolina	No	No	Yes	No	Yes	No	Yes	Yes	Date
South Dakota	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Tennessee	No	Yes	Yes	No	Yes	No	Yes	Yes	Date
Texas	Yes	Yes	Yes	No	No	No	Yes	No	Date
Utah	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Date
Vermont	No	No	Yes	No	Yes	No	No	No	Year
Virginia	Yes	Yes	No	No	Yes	Yes	Yes	No	Date
Washington	Yes	No	Yes	No	Yes	No	Yes	No	Date
West Virginia	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Date
Wisconsin	No	Yes	Yes	No	Yes	No	No	No	No
Wyoming	Yes	Yes	No	No	No	Yes	No	No	No

Table 2: Formats and Available Data in State-level Voter Files

By 2023, nearly all states provided their data in tabular formats, such as CSVs or other delimited files.[3] However, some were stored in relational databases such as .accdb or .mdb files via Microsoft Access (including the most recent files available from Tennessee). Across the time frame, some states radically changed how they reported voter data, while others kept a consistent format. As expected, there is considerable variance in background and demographic information included in the files. Some states provided full dates of birth, as well as information on original and latest registration dates, race, gender, and affiliated party. Other states provided only the year of birth, failing to provide information on the voter's race/ethnicity, partisan affiliation, etc.

A bit of detail is warranted concerning columns 2 and 3 from Table 2. Column 2 is a binary "yes/no" variable that captures whether there is a separate voter history file or not. This information may affect our understanding of who was registered—and eligible to vote—at a given point in time. The significance of this information is evident from the example presented in Tables 3 and 4. The underlying behavior for voters 00001 and 00002 is identical. Table 3 presents their data in what we call a "wide data format," in which a registrant's vote history is not separate from the registration record and is instead provided as additional variables. On the other hand, Table 4 shows their data in what we refer to as a "long data format," in which

[3] North Dakota does not require voters to register to vote.

registration and vote history records are separated, such that additional, detailed information can be obtained.

Voter ID	First Name	Last Name	Reg. Date	Residential State	Gender	Gen. 2022	Gen. 2020	Gen. 2018	...
00001			06/01/2018	GA	M	X	X	X	...
00002			10/01/2020	DC	F		X		...

Table 3: Fictitious Example of a Wide File Format (Vote History and Registration Combined)

Voter ID	Election	Method of Voting	County	Party
00001	Gen. 2018	Polling Place	DeKalb	
00001	Pri. 2018	Early	DeKalb	DEM
00001	Gen. 2020	Mail	Decatur	
00001	Gen. 2022	Mail	Decatur	
00002	Gen. 2018	Polling Place	Clarke	
00002	Pri. 2018	Polling Place	Clarke	REP
...				

Table 4: Fictitious Example of a Long File Format (Separate Vote History File)

Just to be clear, there are two reasons why differences in reporting format might be important. First, as noted above, state voter files using a long data format are more likely to provide information on (a) how the ballots were cast, and (b) the jurisdiction in which the voter cast their ballot. For mail ballots, they might even contain additional information, such as when the ballot was requested, delivered, accepted or rejected, and (if it was rejected) reasons why it was rejected (e.g., "lack of a signature"). Finally, they are also likely to include data on local and state-level turnout history; these data are rarely offered in files arrayed in wide data format. Second, states using a long data format are more likely to keep records of voters even when they have been removed from the voter rolls (most commonly due to voter relocation or death). Although earlier voting records are occasionally deleted, registration lists using long formats are meant to be cumulative and typically contain all records of who was registered and who voted in recent elections. This prevents issues arising from voter file "attrition" following Election Day, which are common and can be problematic for turnout studies (Kim and Fraga 2022). In general, long format files are relatively more likely to help our understanding historical turnout, especially since some wide data files only offer a "last date voted" variable, which is constantly changing for habitual voters.[4]

[4] Indeed, there are ways to make the wide voter file even more uninformative than is shown in Table 3. South Carolina is arguably one of the most uninformative file formats, with column names such as both "General Election Last Voted" and "General Election Previously Voted" in the same voter file snapshot with no intuitive distinctions as to what those dates might be and why might "last" and "previously" be any different.

One downside with long data formats is that they typically do not contain information about registered voters who did not vote. Put another way, eligible nonvoters are much more likely to be missed in these separate vote history files than they are in the wide data (where they are more conspicuous).

The third column in Table 2 shows whether the method of voting is available in the historical files. As we discuss below, some states provide the method of voting, while others provide only a binary "voted/did not vote" variable. Obviously, we prefer the former—more information is always valuable for academics, election officials, and other stakeholders. And we are particularly interested in learning more about whether (or not) registrants are taking advantage of convenience voting methods.

Our review of historical voter files thus yields a straightforward technical recommendation: states ought to provide separate voter "vote history" files, including the method and date of the vote in a specific election. In addition, we recommend that states provide monthly "snapshots" of their voter rolls. These allow researchers and interested observers to "compare and contrast" current voter files with historical records.[5]

2.3 Comparing Raw Voter Files to Administrative Reports of Voter Registration

Given the primacy of turnout as an indicator of the health of American democracy, accurately measuring registration and voter turnout in the U.S. is essential. Thus, another goal of this project is to assess the efficacy of historical voter files towards this end. Accurately estimating the turnout rate means measuring both the number of votes cast (the numerator) and the number of eligible voters (the denominator). Our initial focus has been on the oft-neglected latter number. Prior to conducting a more complex and computationally intensive analysis of the raw voter file data, we sought to evaluate how the number of records in each state's voter file compared to administrative records. We proceed by comparing (1) the statewide counts of registered voters provided by election officials in the U.S. Election Assistance Commission 2020 Election Administration and Voting Survey (EAVS) and (2) statewide counts from voter registration files contemporaneous to the November 2020 election. It is important to note that we selected state-level snapshots of the voter registration lists as close as possible to November 2020. For approximately half of the states, these snapshots reflect the voter registration list as of October 2020. Most of the remaining statewide snapshots are from September, November, or December of 2020. In rare instances, we relied on snapshots from the summer of 2020 or early 2021.

Our method for generating counts of registered voters from the raw voter file data is simple: count the number of rows in the state's voter registration list. For a small number of states, the voter file clearly indicates that a voter has a "canceled" registration or is deceased: we removed these records. Note that this

[5] Wide files can contain information on voting methods by replacing binary turnout with methods of voting utilized by the voter, hence the independence.

does not account for "active" versus "inactive" registrants, duplicate entries, or "deadwood," all of which contribute to discrepancies between official records of the number of voting-eligible residents in a state and the number of registrants (Ansolabehere and Hersh 2014). We prefer the simplicity and objectivity of our measure over more complex modeling, however, as it provides a better test of the viability of raw files as a tool for analyzing voter behavior.

Our method for generating counts from the EAVS is equally simple: we rely on the first question on the survey, which asks election officials to "report the total number of people (not votes or ballots) who were registered and eligible to vote in the November 2020 general election." The EAVS reports results at the sub-state electoral jurisdiction level, generally the county or equivalent sub-state unit responsible for voter registration processing. We sum the values for subpart A of Question A1 across jurisdictions within a state as our baseline indicator of the state-reported number of voting-eligible registrants for the November 2020 election.

In Figure 5, we compare the EAVS estimates to those from the registered voter files. The circles and corresponding percentages indicate the raw file total count divided by the EAVS estimate. For most states, we show appreciable consistency in the registration numbers. In fact, all but six of the states we examine show the raw, unprocessed file count to be between 91% and 102% of the EAVS report. Oregon (118%) and North Carolina (109%) show an "over-estimation" of registrants, while South Carolina (88%), D.C. (84%), Wyoming (81%), and Utah (78%) show an "underestimation."[6]

We also compare registration counts to the 2020 Current Population Survey Voting and Registration Supplement (CPS). As noted by several previous researchers, the CPS's method of establishing registration is the reverse of most political science surveys (e.g., the ANES): respondents are only asked if they are registered to vote *after* stating that they did not vote in the election (Hur and Achen 2013; Pettigrew and Stewart III 2017). We provide CPS estimates of registration totals by state via the gray 95% confidence intervals for these estimates in Figure 5 (Ansolabehere et al. 2022).

As expected, in nearly all cases the CPS offers a much lower estimate of the total number of registrants in a state than the EAVS or the unprocessed historical statewide voter files. Indeed, the average "underestimation" is roughly 30%. This underestimation of the "denominator" is an important (and underrated) reason for the well-documented tendency of surveys, including the CPS, to over-report voter turnout. For 2020, for example, the CPS estimates that 168,307,905 adult citizens were registered to vote (95% CI: 166,937,496 to 169,678,314).

[6] Oregon is a mail-in voting state, which might account for the lower-level of EAVS reported registration. On the other side of the ledger, one explanation for the lower figure in Utah is the ease by which a registrant can opt out of having their registration record provided in the public voter file.

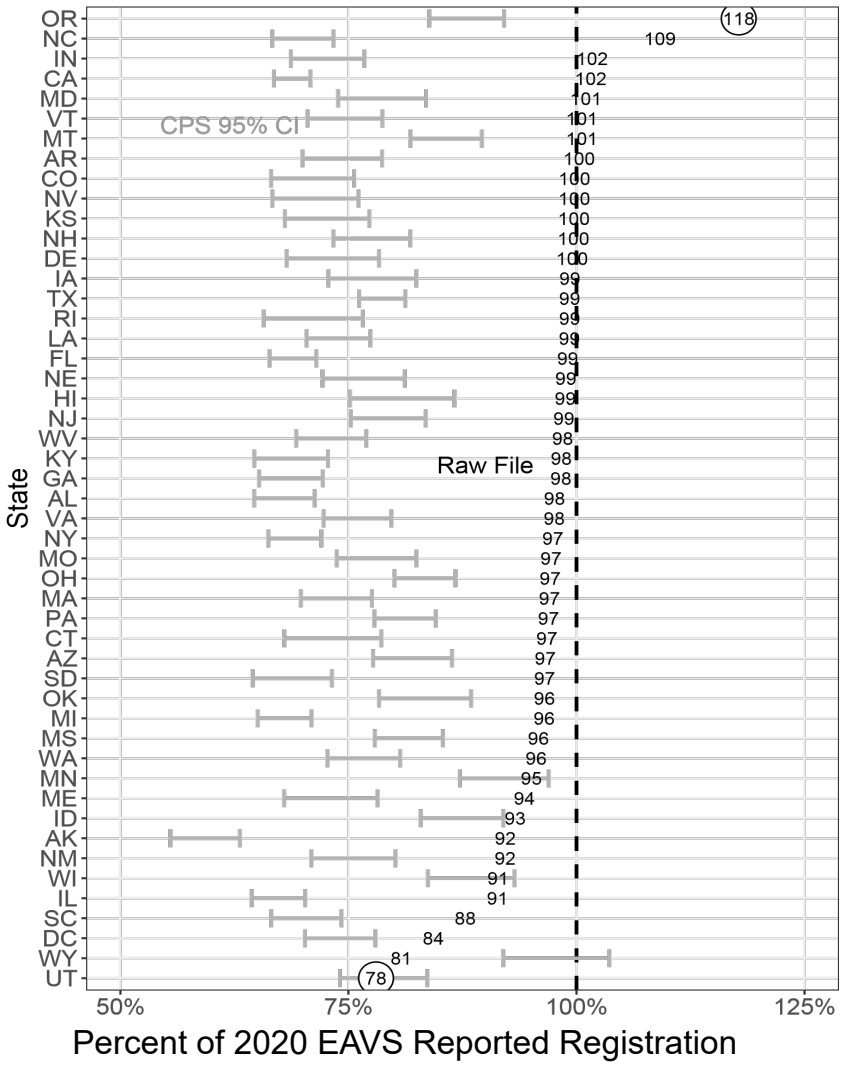


Figure 2: Comparison of Raw Voter Registration File Count to EAVS Reports, 2020.

Notes: The baseline is the sum of county-level EAVS reports for the total number of registered voters (Q.A1a) in the state, represented by a dashed line. Circles and numbers indicate the number of lines in the raw statewide voter registration file closest to the November 2020 election, as a percentage of the EAVS total. Gray error bars indicate a 95% confidence interval for the Current Population Survey estimate of the number of registered voters (PES1 = 1 or PES2 = 1) as a percentage of the EAVS total. North Dakota and Tennessee are not included in the above figure.

In examining the discrepancies between the historical voter files and the EAVS, we explored two explanations for the observed differences. First, we considered the possibility that simple and obvious housekeeping issues were responsible for mismatches in counts of registered voters. For example, among states where the lists have too many registrants, Mississippi and New York have fields identifying "purged" registrants, and Wisconsin defines "inactive" registrants differently than other states. Among states where the lists have too few registrants, California does not list "inactive" registrants. Simply accounting for these issues largely, though not entirely, resolves count discrepancies.

Second, we also considered the possibility that registered voter lists that were relatively distant from Election Day 2020 (say, September 2020 or February 2021) might show more substantial differences when compared to the EAVS data. Although we have not yet finalized this analysis, preliminary estimates indicate that timing is neither systematically nor consistently related to registration count differences (see Figure 3).

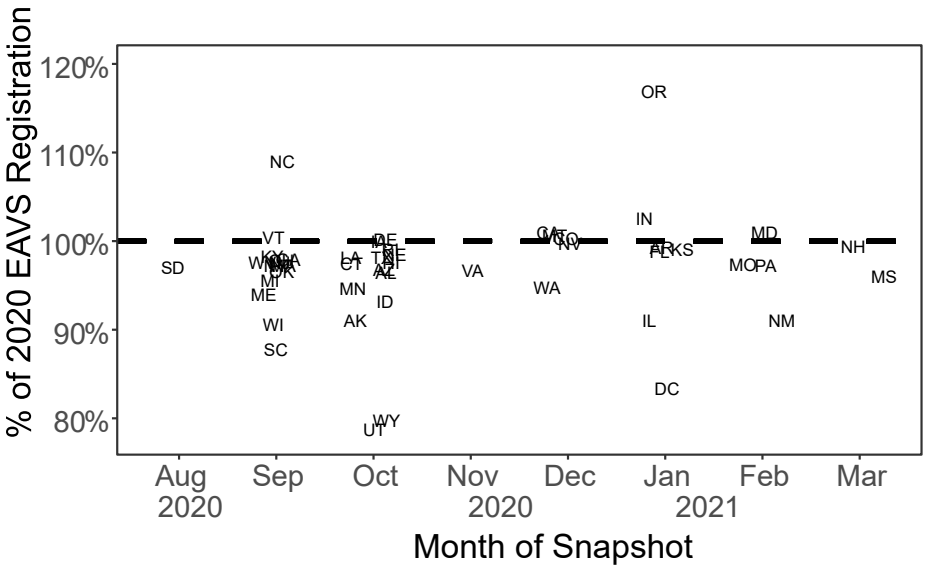


Figure 3: Timing of the Snapshot Closest to the 2020 General Election and the Deviation From the 2020 EAVS Registration

As of now, we have not directly analyzed the efficacy of historical voter files for the measurement of the numerator in the turnout equation: how many registrants voted in an election. While we assume that these records will provide accurate numbers, we think it important to compare these with those derived from surveys and other sources, if only to obtain a better understanding of potential errors in different methods.

As noted above, we are currently cataloging and examining raw historical registered voter files for each state for as many election years as possible. The larger (and more daunting) goal is to create our own individual-level records across elections for registrants from a given state. Of course, current statewide voter records include turnout and (sometimes) method of voting in past elections. But the efficacy of these current records is assumed rather than empirically established. By checking the raw historical records, we can validate or even improve upon existing current records by (a) ensuring that past behavior of registrants is consistently coded, and (b) tracking not only current registrants but those who were registered in a state but either moved out or died. This information is critical information for election administrators as they make resource allocation and policy choices. An important part of our process consists of geocoding registrants by their registration address in each file, which allows us to track them across time and place.

We are currently working on creating replicable codes that can be publicly distributed, which will facilitate processing statewide voter files. We are also working on creating an aggregated set of historical files, which will provide a standardized benchmark for those looking to do overtime analyses of voting behavior.

2.4 Best Practices for Compiling, Disseminating, and Working with Voter Files

In addition to creating (and analyzing) standardized historical statewide voter registration data files provided by L2, over the past several years we have also analyzed a wide range of contemporaneous voter files. Based on this experience, we would recommend the following best practices for compiling, disseminating, and working with voter files.

First, we recommend more analyses of the possible accretion and attrition biases that accompany voter files, especially snap-shot files. There is evidence that some eligible voters are not on official voter lists leading up to Election Day, while nonvoters may be inadvertently removed after Election Day (Kim and Fraga 2022). This might affect the number of registered voters (if not the recorded votes), although there is little evidence of a partisan bias to this phenomenon. But there is much we do not know about this.

Second, and consistent with our earlier recommendation, we urge state officials to provide—and researchers to leverage—snap-shot lists of registered voters. Recall that in Figure 3, we see that there did not seem to be a clear pattern between the timing of the list and registration numbers for the 2020 election, but this is only a single data point. For other variables, snapshot timing may have an impact. Consider the analysis of the so-called "partisan gap" between voters and nonvoters. Let us define the partisan gap as the proportion of Democrats among nonvoters minus the proportion of Democrats among voters. In this set-up, a positive number indicates that nonvoters are disproportionately Democrats compared to voters, whereas a negative number indicates the opposite. In Figure **4** we look at the partisan gap in North Carolina to see how the timing of the voter list correlates with the estimate. The data indicate that the longer one waits to procure the data after the election, the more likely you are to underestimate the partisan gap.[7]

[7] This even though North Carolina provides separate voter history data that we earlier identified as a "best practice."

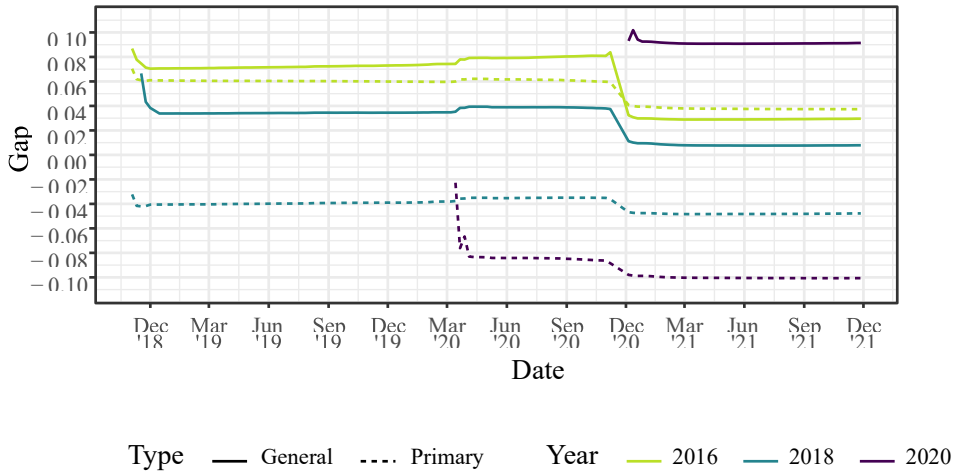


Figure 4: Partisan Gap between Voters and Non-voters, North Carolina

3 Evaluating Changes in Vote Mode, Aggregate versus Individual-Level

3.1 General Approach

Throughout this report, we have emphasized that the methods (or modes) by which people vote is an increasingly important part of political behavior. The MIT Data Lab estimates that so-called "convenience" voting increased from approximately 15% in 2000 to 73% in 2020 (https://electionlab.mit.edu/research/voting-mail-and-absentee-voting). The use of mail voting in American politics has been rapidly accelerating, and amid the COVID-19 pandemic, it was estimated that 46% voted by absentee or mail-in ballot (Doherty et al. 2020). Changes over time in how people choose to vote are especially important not only because of new convenience voting measures, but also because of the shock of the COVID-19 pandemic that undoubtedly disturbed existing voting habits.

Although there are numerous questions one might ask about trends in the method of voting, we are particularly interested in ascertaining the extent to which vote mode varies by individual voter traits. That is, are certain types of voters more likely to vote by absentee/mail ballot? Are younger voters, or racial and ethnic minorities more likely to prefer this method of voting? If so, then perhaps some appreciable portion of the increase in convenience voting is driven by the changes in the composition of voters.

The historical statewide voter files offer a unique possibility to analyze how it is that voters are casting their ballots and the extent to which *individuals are changing* their method of voting over time. The intriguing possibility—on that has not found much support in previous studies—is that nonvoters may be becoming voters as additional methods of casting a ballot are offered. Looking at statewide voter files from 2012-22, our analysis offers insight into the suitability (and limits) of historical voting records for engaging such a question. We then estimate the extent to which changes in the use of different voting modes are driven by (1) the addition of new voters, (2) deletion of peripheral voters, or (3) changes in the behavior of existing voters. We also analyze heterogeneity by geography and the legal environment.

3.2 Limitations of the Data

Despite our favorable assessment of the overall potential of historical voter file data, when it comes to voting method, it is important to reiterate that relevant information does not exist across the board. Among the fifty states, only thirty states provide data on how voters cast their ballots (D.C. does as well). Twenty states report only whether a registrant voted.[8]

Of course, states that do not report voting method may not offer easily available absentee/mail-in ballots or an in-person early voting option. Alabama, for instance, does not offer "no-excuse" absentee voting, and a set of criteria must be met before voters can apply for an absentee ballot. This application must then be pre-approved, and the mail-in ballot itself must be delivered before noon on Election Day. Our view, however, is that information on who uses these methods (restricted or not) is still valuable. Indeed, such information may shed light on how different voting methods impact turnout across the states. Furthermore, we may learn much about voting by analyzing voters who choose relatively "difficult" voting methods, such as mail-in voting in Alabama or in-person voting in mail-in state like Oregon (Kim et al. 2022). Finally, tracking and analyzing voters' method of casting a ballot may speak to broader topics, such as trust in institutions, partisan polarization, and more.

Still, while we may be able to estimate how much absentee voting occurred in the twenty states based on statewide summaries[9], we will not be able to analyze changes in voting method or the adoption of new voting methods at the individual-level. This is not insignificant, as there may be slight differences in partisan leanings of the states that disclose methods of voting and those that do not, which means we cannot assume our missing data are randomly distributed. In fact, a cursory glance at the aggregate-level data

[8] Alabama, Hawaii, Delaware, Illinois, Kansas, Kentucky, Louisiana, Massachusetts, Michigan, Missouri, Mississippi, Montana, North Dakota, Nebraska, Ohio, Oregon, Rhode Island, South Carolina, Vermont, and Washington only report turnout.

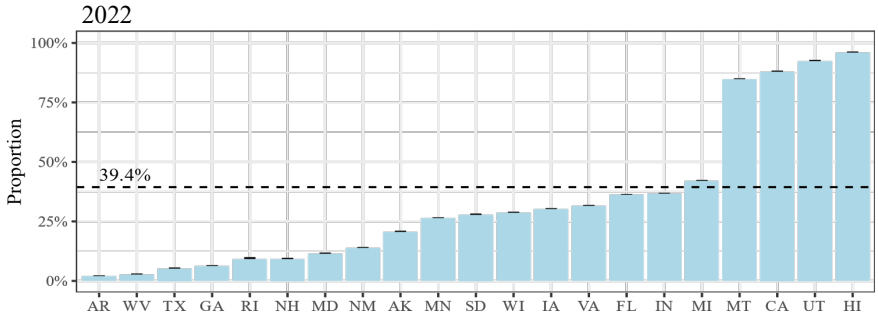
[9] For example, see https://www.ohiosos.gov/media-center/press-releases/2022/2022-11-07/.

confirms the face-validity of this possibility: Biden won in 54.8% of the two-party presidential vote in 2020 in states where the voting method is disclosed, while he won only 45.0% elsewhere.

3.3 Descriptive Statistics

Figure 5 shows the proportion of mail-in voting utilized in the past three general elections—2018, 2020, and 2022—in the thirty states where the data are available.[10] It seems likely that the absence of the twenty states that do not keep records of voting method slightly inflates our estimated percentages. Still, our numbers are like those estimated by well-regarded surveys. For example, our estimate of mail-in voting in 2020 (51.5%) is only five points higher than that of the Pew Research report Doherty et al. (2020).

We would also observe that over half of the vote in 2020 was by mail. In historical terms, this is a stunning number. Clearly, the COVID-19 pandemic and the attendant de-emphasis of in-person voting accelerated the existing trend towards increased convenience voting. Still, it is also true there is a 9-point increase in mail-in voting from the 2018 midterm election (before COVID) to the 2022 midterms (mostly after COVID).[11] Given these results, one might reasonably surmise that exposure to a more permissive absentee/mail-in voting environment allowed many voters to explore voting methods they otherwise would not have tried, which led to some appreciable portion of these voters switching their preferred voting method.



[10] Some of the states are not displayed because their data wrangling warrants a conversation with election administrators about their accuracy. The figures will be edited in a future version of the report to include more states.
[11] Officially, the World Health Organization (WHO) declared an end to the global Public Health Emergency (PHE) for COVID-19 on May 3, 2023.

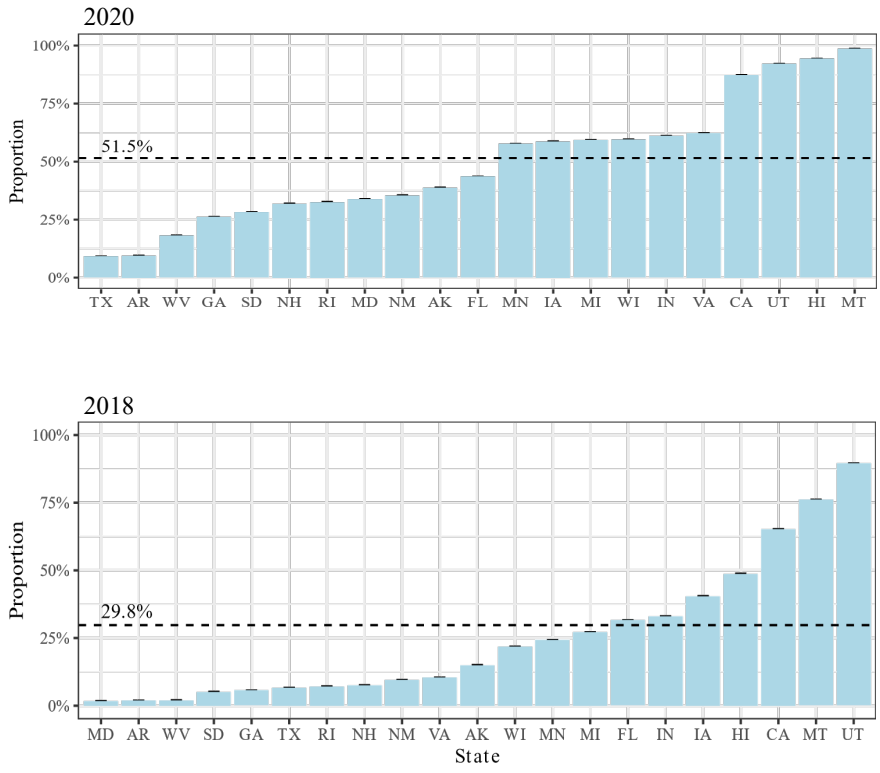


Figure 5: Proportion of Mail Voting Nationwide and Across States

3.4 Changes in Voting Methods Over Time: Composition or Changing Choice?

But a reasonable surmise is not a definitive answer. While it is possible that voters are changing their preferred mode of voting, it is also possible that people who voted in 2018 are simply different than people who voted in 2020 or 2022. In other words, we cannot ascertain from our descriptive data or from existing survey data whether increased absentee/mail-in voting is mostly due to "core" voters *switching* to mail-in balloting or whether it is mostly due to *new voters* who prefer to vote by mail.

What happens if we ignore the possibility that increased mail-in voting is primarily driven by changes in the composition of the electorate? Consider the implied rate of switching to mail-in voting in Georgia if we ignore shifts in who votes. According to the secretary of state's office, 5.6% of Georgia voters voted absentee-by-mail in the November 2018 election; this number nearly quintuples to 26.0% in November 2020; it then drops down to 6.2% in November 2022. The (naïve) implication is that roughly 20% of Georgians shifted to vote by mail *and then shifted back to in-person methods* in the span of three election cycles. However, our historical voter records—which amount to panel data tracking individual-level behavior—show that only 11.9% of 2022 voters followed that pattern. This percentage is smaller than the share voting for the first time in the 2022 election (13.7%). It is also smaller than the share that voted

in 2020 and 2022 but not 2018 (19.3%), or the share that switched between voting early in-person and voting on Election Day (or vice versa) in 2020 and 2022 (16.7%). Perhaps surprisingly, this last percentage is *larger* than the number shifting from mail to an in-person method between 2020 and 2022 (16.4%), indicating that the largest share of method "switchers" in 2022 were in-person voters rather than mail voters from 2020.

The upshot is that a naive estimate of switching—the difference between the proportion of mail voting between elections—can be significantly different from the actual switching rates. Table 5 shows the proportion of absentee/mail-in mail voters in each of the last three general elections, while also gauging election-to-election consistency (or inconsistency) in voting method by sorting behavior into four categories: (1) in-person to in-person, (2) in-person to mail, (3) mail to mail, and (4) mail to in-person.

State	% Mail 2018	% Mail 2020	% Mail 2022	In-person 2018, In-person 2020	In-person 2018, Mail 2020	Mail 2018, Mail 2020	Mail 2018, In-person 2020	In-person 2020, In-person 2022	In-person 2020, Mail 2022	Mail 2020, Mail 2022	Mail 2020, In-person 2022	In-person 2018, In-person 2022	In-person 2018, Mail 2022	Mail 2018, Mail 2022	Mail 2018, In-person 2022
AK	14.9%	38.7%	20.5%	49.6%	23.9%	9.3%	3.3%	37.6%	3.1%	10.7%	14.2%	50.4%	8.8%	5.4%	4.2%
AR	1.8%	9.4%	1.9%	80.2%	7.6%	0.9%	0.4%	55.9%	0.3%	0.8%	4.4%	66.2%	0.9%	0.3%	0.4%
CA	64.6%	86.8%	87.8%	8.2%	24.9%	59.3%	2.5%	3.7%	3.2%	51.7%	3.6%	6.1%	17.4%	48.0%	2.2%
FL	17.1%	31.7%	18.1%	45.1%	17.4%	24.0%	4.2%	31.8%	2.2%	21.3%	6.8%	36.3%	9.0%	17.2%	4.5%
GA	3.6%	21.0%	3.9%	65.5%	19.8%	3.5%	1.3%	52.7%	0.5%	4.0%	13.9%	68.3%	3.0%	1.5%	2.1%
HI	48.7%	95.2%	96.0%	0.5%	6.1%	43.5%	0.5%	0.4%	0.6%	26.4%	0.5%	0.0%	0.0%	0.0%	0.0%
IN	32.9%	60.9%	36.4%	24.3%	25.5%	22.3%	2.4%	15.4%	1.7%	16.9%	15.8%	24.5%	7.2%	11.7%	5.4%
IA	40.4%	58.9%	30.1%	29.3%	25.2%	31.9%	4.2%	24.9%	1.5%	19.3%	21.9%	37.4%	6.5%	17.3%	11.6%
MI	26.8%	59.3%	42.1%	31.8%	31.9%	22.3%	1.3%	26.9%	2.1%	30.0%	15.0%	38.6%	16.0%	17.7%	2.4%
MN	24.2%	57.6%	26.3%	34.2%	35.6%	19.7%	2.3%	27.3%	1.7%	17.2%	24.9%	48.5%	9.2%	11.5%	6.9%
MT	76.0%	98.6%	84.6%	1.3%	21.9%	70.4%	0.1%	0.9%	0.2%	66.4%	11.3%	10.6%	7.1%	58.8%	1.8%
NH	7.5%	31.8%	9.2%	58.7%	26.0%	4.5%	2.0%	45.0%	1.8%	4.6%	17.2%	65.3%	4.9%	2.2%	2.9%
NM	9.4%	35.4%	13.8%	55.0%	26.4%	6.2%	1.7%	43.0%	1.2%	8.3%	15.7%	60.2%	6.0%	3.5%	2.5%
RI	7.0%	32.5%	9.3%	58.4%	25.4%	3.6%	2.1%	42.2%	0.8%	5.0%	14.9%	60.8%	4.8%	1.4%	2.8%
SD	26.4%	51.3%	28.2%	38.8%	26.4%	20.0%	3.0%	32.6%	2.6%	18.3%	18.8%	46.7%	9.0%	13.0%	5.9%
TX	6.6%	9.1%	5.2%	78.5%	5.2%	3.9%	1.5%	58.4%	0.6%	2.8%	2.9%	64.1%	1.8%	2.2%	1.8%
UT	89.3%	92.0%	92.4%	1.5%	7.7%	80.6%	2.7%	1.3%	2.2%	59.3%	3.0%	1.5%	4.7%	64.6%	2.8%
VA	9.9%	62.6%	32.0%	30.7%	50.2%	7.5%	0.9%	19.9%	1.7%	18.5%	22.6%	43.3%	18.1%	4.3%	2.2%
WV	2.0%	18.1%	2.6%	72.0%	15.7%	1.1%	0.3%	46.7%	0.3%	1.2%	8.3%	62.1%	1.3%	0.4%	0.3%
WI	12.9%	41.1%	15.6%	29.2%	36.8%	16.3%	1.7%	25.1%	1.5%	19.2%	24.1%	41.9%	10.9%	9.5%	5.0%

Table 5: Proportion of Mail Voting and Shifts in Methods of Voting, 2018–2022 General Elections

Table 6 further simplifies this information into implied vs. actual switching behavior for a succinct comparison. For both tables, fields such as "In-person 2018, In-person 2020" indicate the percentage of voters who voted in-person in 2018 and in 2020, against the number of voters in 2018. Tables 5 and 6 clearly demonstrate that the changes in the percentage of mail voting is a problematic way to ascertain whether existing voters are switching (or not switching) their method of voting. An archetypical example of this is Florida, where the "implied" switching rate is 13.6%p, but only 6.8% of 2018 voters switched.

State	Implied Switching to Mail, 2020	In-person 2018, Mail 2020	Implied Switching to In-person, 2022	Mail 2020, In-person 2022
AK	23.8%p	23.9%	18.1%p	14.2%
AR	7.6%p	7.6%	7.5%p	4.4%
CA	22.2%p	24.9%	-1.0%p	3.6%
FL	14.6%p	17.4%	13.6%p	6.8%
GA	17.4%p	19.8%	17.2%p	13.9%
HI	46.4%p	6.1%	-0.9%p	0.5%
IN	27.9%p	25.5%	24.4%p	15.8%
IA	18.5%p	25.2%	28.8%p	21.9%
MI	32.4%p	31.9%	17.2%p	15.0%
MN	33.4%p	35.6%	31.3%p	24.9%
MT	22.6%p	21.9%	14.0%p	11.3%

NH	24.4%p	26.0%	22.6%p	17.2%
NM	26.0%p	26.4%	21.7%p	15.7%
RI	25.5%p	25.4%	23.2%p	14.9%
SD	24.9%p	26.4%	23.1%p	18.8%
TX	2.5%p	5.2%	3.9%p	2.9%
UT	2.8%p	7.7%	-0.3%p	3.0%
VA	52.7%p	50.2%	30.6%p	22.6%
WV	16.2%p	15.7%	15.5%p	8.3%
WI	28.1%p	36.8%	25.4%p	24.1%

Table 6: Implied Switching vs. Actual Switching of Voting Methods, 2018–2022 General Elections

Table 7 provides another take on this issue by showing conditional probabilities based on past choices. In California, for example, conditional on being an in-person voter in 2018 general election, 70.4% voted by mail in the 2020 election. This is not surprising, given that the 2020 election was the first in which California sent every registered voter a postage pre-paid absentee ballot.

State	In-person 2018, In-person 2020	In-person 2018, Mail 2020	Mail 2018, Mail 2020	Mail 2018, In-person 2020	In-person 2020, In-person 2022	In-person 2020, Mail 2022	Mail 2020, Mail 2022	Mail 2020, In-person 2022	In-person 2018, In-person 2022	In-person 2018, Mail 2022	Mail 2018, Mail 2022	Mail 2018, In-person 2022
AK	58.3%	28.1%	62.3%	22.0%	61.2%	5.0%	27.8%	36.6%	59.3%	10.3%	36.3%	28.5%
AR	81.6%	7.8%	50.2%	20.1%	61.7%	0.3%	8.8%	46.6%	67.5%	0.9%	18.8%	21.6%
CA	23.1%	70.4%	91.8%	3.9%	27.9%	24.4%	59.6%	4.2%	17.1%	49.1%	74.3%	3.5%
FL	65.8%	25.4%	76.1%	13.3%	56.2%	3.8%	49.0%	15.6%	53.0%	13.1%	54.5%	14.3%
GA	69.5%	21.0%	60.9%	21.9%	71.7%	0.7%	14.9%	52.6%	72.6%	3.1%	25.6%	36.1%
HI	7.0%	84.2%	89.3%	1.0%	7.6%	11.9%	27.8%	0.6%	0.0%	0.0%	0.0%	0.0%
IN	36.3%	38.0%	67.6%	7.3%	39.3%	4.3%	27.7%	25.9%	36.5%	10.8%	35.7%	16.3%
IA	49.2%	42.3%	79.1%	10.5%	60.4%	3.7%	32.8%	37.3%	62.7%	10.9%	42.7%	28.7%
MI	43.5%	43.6%	83.2%	5.0%	66.0%	5.1%	50.6%	25.3%	52.8%	21.9%	66.1%	9.1%
MN	45.2%	46.9%	81.4%	9.6%	64.5%	4.0%	29.9%	43.3%	64.0%	12.1%	47.4%	28.3%
MT	5.2%	91.1%	92.6%	0.2%	61.4%	10.5%	67.4%	11.5%	43.9%	29.5%	77.5%	2.4%
NH	63.4%	28.1%	60.8%	26.5%	66.1%	2.6%	14.5%	54.0%	70.5%	5.3%	29.0%	38.2%
NM	60.7%	29.2%	66.2%	18.4%	66.6%	1.8%	23.4%	44.5%	66.5%	6.6%	36.9%	26.6%
RI	62.8%	27.4%	50.7%	30.1%	62.5%	1.2%	15.4%	45.9%	65.4%	5.2%	20.0%	40.2%
SD	52.8%	35.9%	75.5%	11.2%	67.1%	5.3%	35.7%	36.4%	63.5%	12.2%	49.2%	22.4%
TX	84.0%	5.5%	60.0%	22.2%	64.2%	0.7%	31.3%	31.4%	68.6%	1.9%	34.1%	28.2%
UT	13.8%	72.2%	90.3%	3.0%	16.4%	27.6%	64.4%	3.3%	13.7%	44.0%	72.4%	3.1%
VA	34.1%	55.8%	75.5%	8.8%	53.2%	4.5%	29.5%	36.2%	48.1%	20.1%	43.7%	22.4%
WV	73.4%	16.0%	54.4%	14.5%	57.0%	0.3%	6.8%	45.9%	63.4%	1.3%	18.0%	15.4%
WI	37.2%	46.8%	75.9%	7.8%	62.4%	3.6%	32.0%	40.4%	53.4%	13.8%	44.5%	23.1%

Table 7: Switching Voting Methods of voting Conditional on Past Choices, 2018–2022 General Elections

In states without significant systematic changes, however, the conditional probabilities establish that a majority of those who voted in person in 2018, did so in 2022 as well.[12] Similarly, more 2018 absentee/mail-in voters opted to vote by mail than in-person in 2022, although there was substantial attrition (due to nonvoting). It is therefore reasonable to conclude that, *ceteris paribus*, voters tend to stick with their preferred method of voting. This finding is consistent with recent studies that show a voter's preferred method for casting a ballot seems to be habitual (Kim et al. 2022). More research is necessary to assess what causes switchers to do so, as well as how initial habits of preferred voting methods form.

[12] Because we account for nonvoters, the sum of columns that display conditional probabilities for a given election may not equal 100%.

3.5 Missing Data and Possibilities with More Data

Some further discussion of missing data and inconsistent reporting is warranted. It is possible that heterogeneity in data reporting may be affecting the descriptive data offered in this report. In particular, the use of absentee/mail-in ballot drop-boxes and polling place drop-offs can affect whether a ballot is counted as "absentee." In other words, the absence of more detailed data on the delivery of the ballot may skew our understanding of voting method.

Of course, even if the ballot delivery is reported in detail, the classification of voting methods is not straightforward. For example, how should we treat voters who have requested absentee ballots but have chosen to drop off their ballots at a vote center or polling place? Consider Hawaii, which is now an all-mail voting state. The Aloha state records whether a voter casts their ballot by mail or at a walk-in polling place.[13] In 2018, when polling places were still utilized, 15.0% of voters who requested absentee ballots had dropped their ballot off at a polling place. Even in 2020 and 2022, as the pandemic raged on, 4.8% and 4.0% of voters (respectively) walked in with the absentee ballot. Should we count these voters as "in-person," "mail-in," or some create some classificatory hybrid?[14] These sorts of questions should not be interpreted as suggesting that we do not need (or cannot handle) detailed data on ballot delivery. Rather, we think adding one or two additional pieces of information to statewide voter files would significantly enhance our ability to do high-level, transparent, apples-to-apples comparisons.

In thinking through knotty questions like these, we also recognize the importance of establishing partnerships with election administrators. These officials might aid us in two specific regards:

- They might help us procure more detailed data not available on from publicly disbursed voting records,
- They might clarify what certain voting method classifications mean in practice,
- They might elucidate how timing (when a ballot is received) affects the treatment and classification of ballots.

In addition, the analyses we conduct might help election officials to:

- Allocate their resources more effectively across different voting methods,
- Allocate their resources more effectively across the election/voting calendar.

[13] This is recorded as "AB Walk" in the voter history records.
[14] In this report, we classify these voters as "in-person" voters. We also code as "in-person" voters for whom we have duplicate records of absentee and in-person voting.

4 Products

The following working papers and presentations rely on data from this project:

- SPSA 2023 Presentation: *Leveraging Historical Voter Files as Accurate Measures of Who Votes*
- ESRA 2023 Presentation: *Disaggregating Choices and Changes in Methods of Voting*
- Presentation at the Brennan Center for Justice, 2023: *Extending Ecological Inference with Vote Mode Data*
- Working paper: *When Do Voter Files Accurately Measure Turnout? How Transitory Voter File Snapshots Impact Research and Representation*
- SPSA 2024 Presentation (scheduled): *Switchers or New Registrants? Analyzing Trends in Vote Method Usage*

5 Participants

Beyond the principals and graduate research assistants involved in our project (e.g., Cornelia Lawrence at UT-Austin), we have also been in contact with Prof. Paul Gronke (Reed College) and Prof. Mindy Romero (USC)—who are conducting surveys in states to assess ballot tracking—about our historical voter files. In addition, we have also been contacted by Stephanie Puello (Ph.D. candidate, UC-Denver) about using our files for her research into statewide voter restoration practices.

6 Impact

Thus far, this project has focused on acquiring historical voter files and rendering the attendant data in a coherent and standardized manner for researchers. As such, its impact will be on future research. More specifically, despite limitations across states and over time, we are optimistic that this project is a promising first step in recreating and validating a turnout record for much of the present century. This will facilitate empirically informed assessments of shifts in who votes, how many voters are voting, where turnout lags and where it leads, and how voters are casting their ballots. These assessments are critical to a fuller understanding of democratic functioning in the U.S.

In more immediate terms, the data are already yielding intriguing results on more focused questions. Of note is our finding that most of the changes in voting methods are driven by new and different voters casting ballots, and that few voters change their method of voting from election to election. Disrupting traditional methods of voting may, therefore, have negative consequences, especially for "regular" voters accustomed to casting their ballots in a particular way.

While not the subject of this current report, we would also note that we have conducted preliminary analyses showing that "deadwood" is much less common in historical statewide voter registration records than is sometimes claimed. Should these results hold, this finding ought to inform standards and best practices regarding the nature and timing of voter roll purges.

7 Problems, Changes, and Next Steps

A main goal of this project is to develop a comprehensive historical voter registration data set that could be used by election officials and academic researchers. This remains the goal, although we have become aware of limits with respect to these historical records. More specifically, we have encountered issues regarding how to make these records easily available to other researchers. Our agreement with L2 allows us–the researchers who signed the contract–to access, clean, and analyze the data. However, we cannot make the underlying data available to other researchers without explicit permission from L2. The process by which someone outside the team would apply and be granted access continues to be a work in progress.[15]

As noted throughout the report, some concerns about the data remain. Some states have more complete and robust records going back further in time. More recent years are, therefore, in better shape than more distant years. A good example is that our records concerning methods of voting are better in 2020-22. Even on this measure, though, there are states that do not record method of voting. Obviously, this will require extrapolation strategies that reduce the precision of our estimates. While this is occasionally frustrating, it is important to know.

Our immediate next step is to continue to develop and compile replicable code. While many working hours from the PIs and research assistants went into wrangling the raw voter files across the 51 jurisdictions, there are still many things to clean, code, and verify before sharing the data more broadly. We hope to achieve this in the next academic year.

Analytically, we hope to expand on our analyses of voter registration and voter turnout so that we can offer a more comprehensive analysis of the relative efficacy of voter files for estimating aggregate- and individual-level turnout rates. We also intend to continue our examination of "deadwood" within these historical voter files, and to refine and expand our analysis of the use of different voting methods.

[15] One anonymous reviewer of this report chastised the PIs for not making public accessibility a condition of the data acquisition. While we appreciate this critique, our grant proposal clearly stated that this was something we would *try to negotiate* with L2 in conjunction with MEDSL (who administered the grant). Put succinctly, unlimited access to privately held data with any commercial value is unlikely. Of course, we might have been able to obtain exclusive rights to the underlying data had we offered overwhelming compensation. The rather hurried timing of this round of grants further constrained us. Still, we continue to work with L2 to make these data available to scholars on a case-by-case basis.

References

Ansolabehere, Stephen, Bernard L. Fraga, and Brian F. Schaffner (2022). The current population survey voting, and registration supplement overstates minority turnout. *Journal of Politics 84*(3), 1850–1855.

Ansolabehere, Stephen and Eitan Hersh (2012). Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. *Political Analysis 20*(4), 437–459.

Ansolabehere, Stephen and Eitan Hersh (2014). Voter Registration: The Process and Quality of Lists. In B. C. Burden and C. Stewart, III (Eds.), *The Measure of American Elections*. Cambridge: Cambridge University Press.

Ansolabehere, Stephen, and Nathaniel Persily (2010). Measuring Election System Performance. *New York University Journal of Legislation and Public Policy 13*(3), 445–470.

Berent, Matthew K., Jon A. Krosnick, and Arthur Lupia (2016). Measuring Voter Registration and Turnout in Surveys. *Public Opinion Quarterly 80*(3), 597–621.

Burden, Barry C. (2000). Voter Turnout and the National Election Studies. *Political Analysis 8*(4), 389–398.

Doherty, Carroll, Jocelyn Kiley, Nida Asheer, and Calvin Jordan (2020). Sharp Divisions on Vote Counts, as Biden Gets High Marks for His Post-Election Conduct. Technical report, Pew Research Center.

Enos, Ryan D., Anthony Fowler, and Lynn Vavreck (2014). Increasing Inequality: The Effect of GOTV Mobilization on the Composition of the Electorate. *The Journal of Politics 76*(1), 273–288.

Fraga, Bernard L. (2018). *The Turnout Gap: Race, Ethnicity, and Political Inequality in a Diversifying America*. Cambridge University Press.

Green, Donald P., and Alan S. Gerber (2005). Recent Advances in the Science of Voter Mobilization. *The ANNALS of the American Academy of Political and Social Science 601*(1), 6–9.

Grimmer, Justin, Michael C. Herron, and Matthew Tyler (2023). Evaluating a New Generation of Expansive Claims about Vote Manipulation.

Hersh, Eitan D. (2015). *Hacking the Electorate: How Campaigns Perceive Voters*. Cambridge University Press.

Hur, Aram, and Christopher H. Achen (2013). Coding Voter Turnout Responses in the Current Population Survey. *Public Opinion Quarterly 77*(4), 985–993.

Igielnik, Ruth, Scott Keeter, Courtney Kennedy, and Bradley Spahn (2018). Commercial Voter Files and the Study of U.S. Politics. Technical report, Pew Research Center.

Jackman, Simon, and Bradley Spahn (2019). Why Does the American National Election Study Overestimate Voter Turnout? *Political Analysis 27*(2), 193–207.

Kim, Seo-young Silvia, Akhil Bandreddi, and R. Michael Alvarez (2022). Why Vote in Person in a Pandemic? Using Machine Learning to Predict Voting Methods.

Kim, Seo-young Silvia, and Bernard Fraga (2022). When Do Voter Files Accurately Measure Turnout? How Transitory Voter File Snapshots Impact Research and Representation.

McDonald, Michael P. (2007). The True Electorate: A Cross-Validation of Voter Registration Files and Election Survey Demographics. *Public Opinion Quarterly 71*(4), 588–602.

Pettigrew, Stephen, and Charles Stewart III (2017). Moved out, moved on: Assessing the effectiveness of voter registration list maintenance.

Wickham, Hadley (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software 40*(1), 1–29.

8 Appendix: Sankey Plots of Vote Mode Change by State, 2020–2022

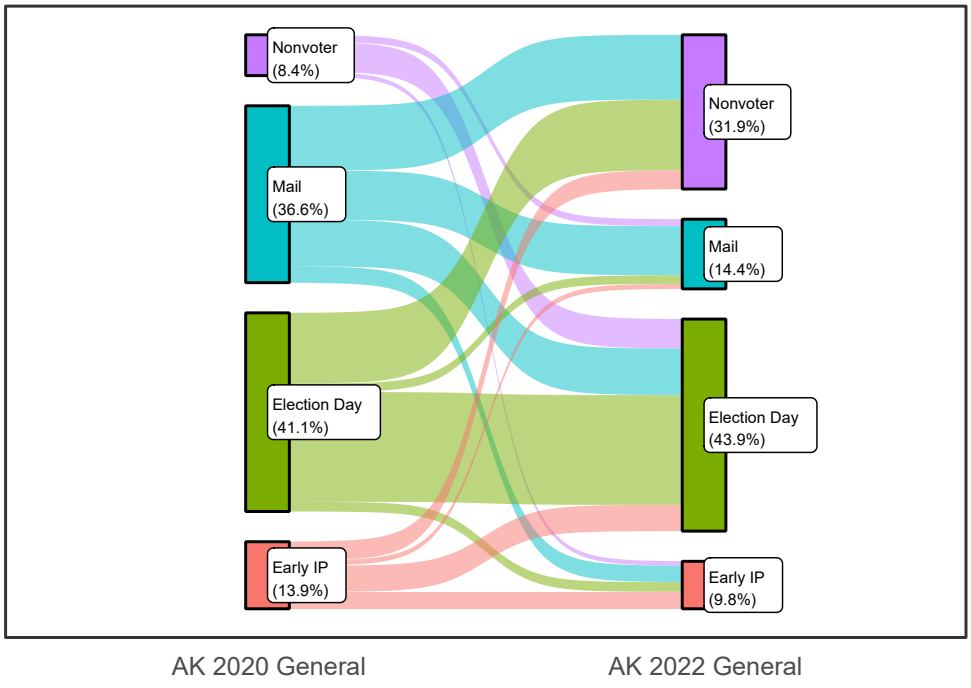


Figure 6: AK Actual Vote Mode Shift for Individual Voters, 2020-2022

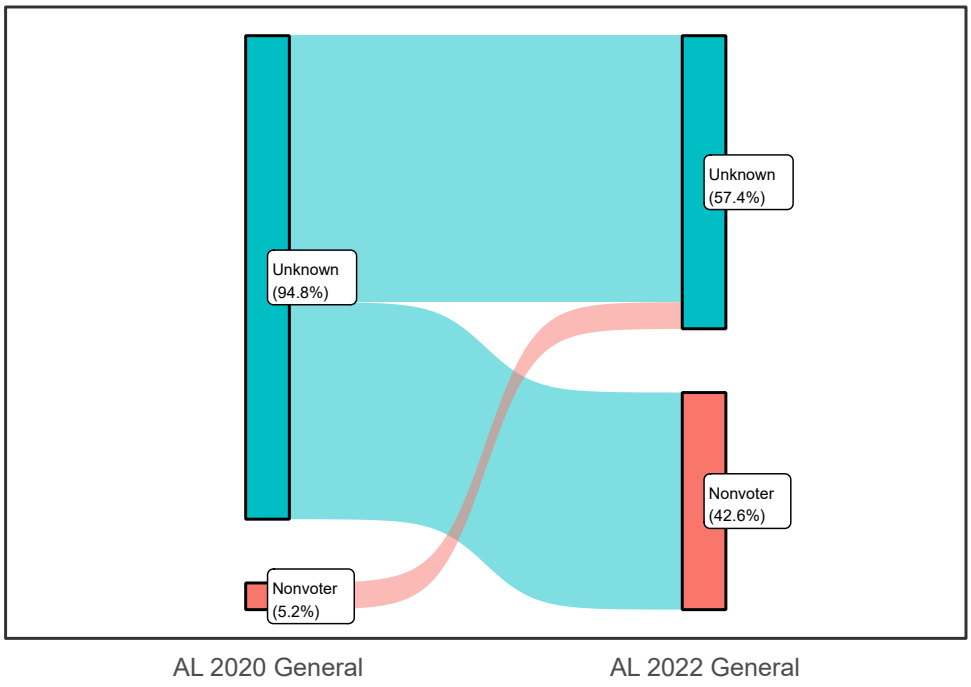


Figure 7: AL Actual Vote Mode Shift for Individual Voters, 2020-2022

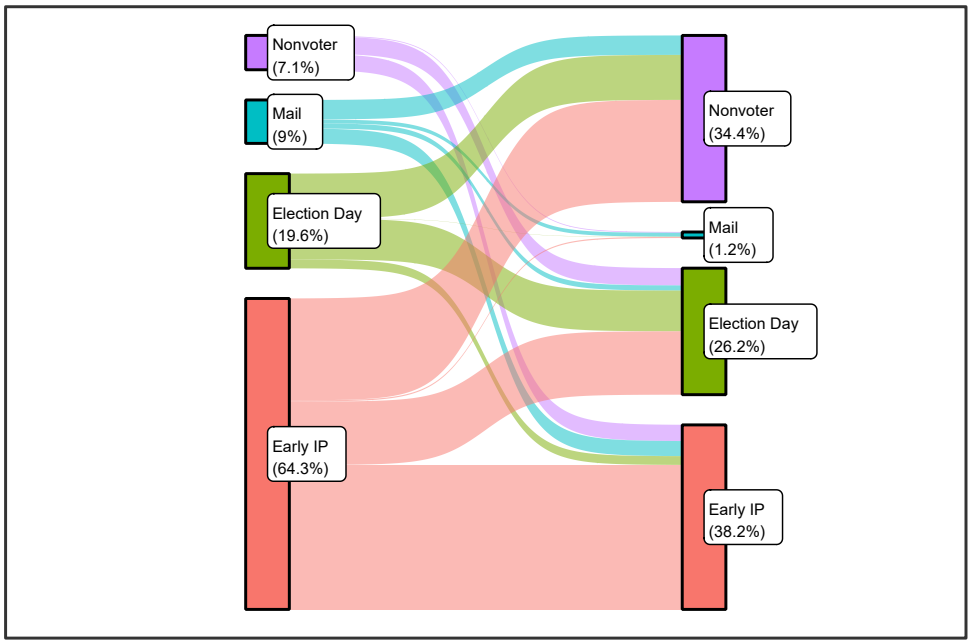


Figure 8: AR Actual Vote Mode Shift for Individual Voters, 2020-2022

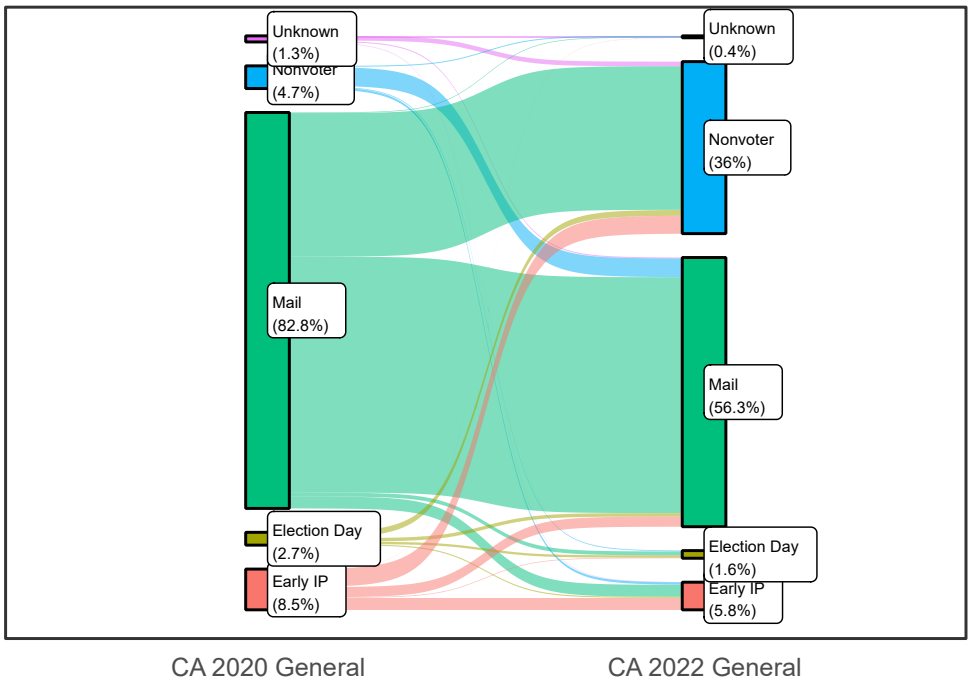


Figure 9: CA Actual Vote Mode Shift for Individual Voters, 2020-2022

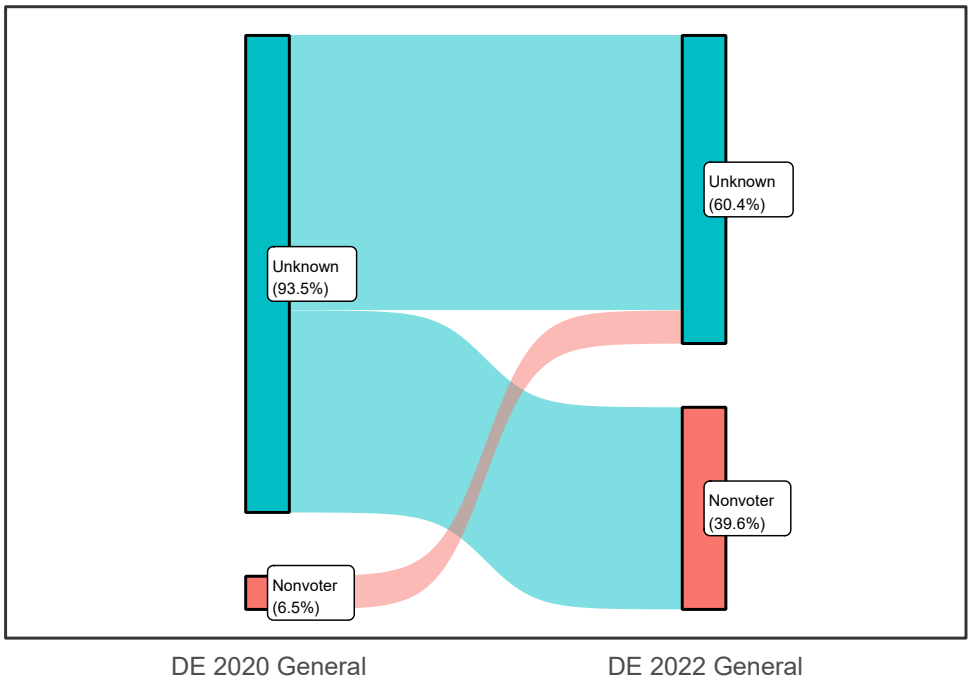


Figure 10: DE Actual Vote Mode Shift for Individual Voters, 2020-2022

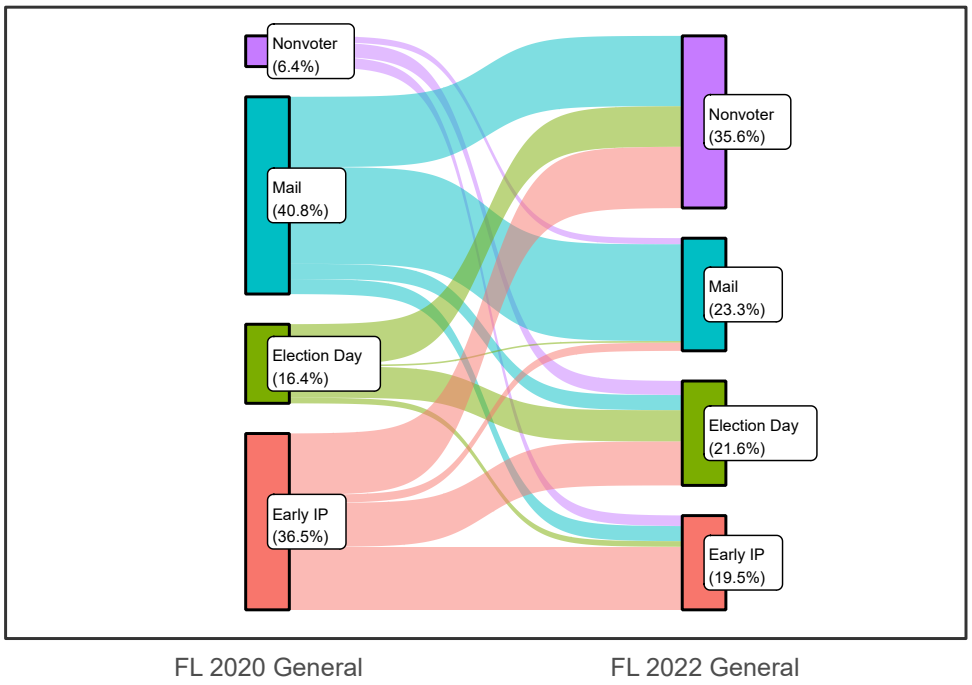


Figure 11: FL Actual Vote Mode Shift for Individual Voters, 2020-2022

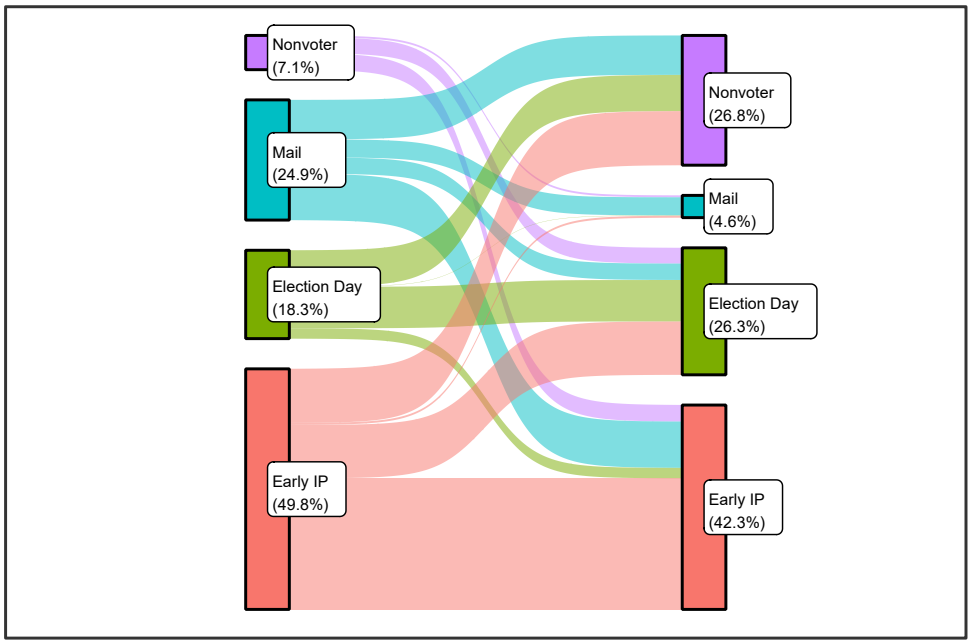


Figure 12: GA Actual Vote Mode Shift for Individual Voters, 2020-2022

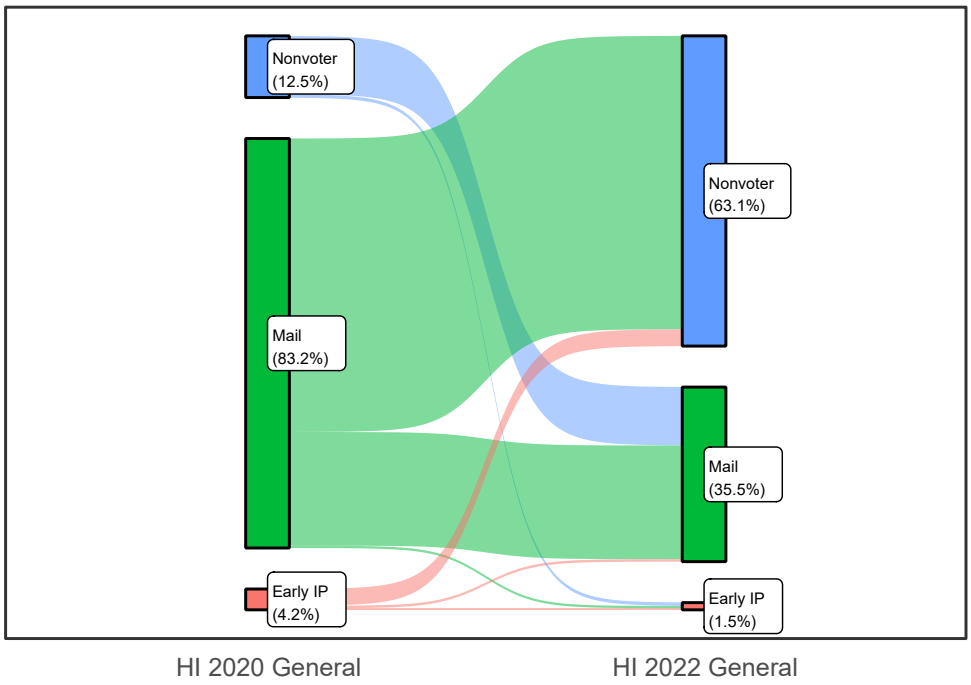


Figure 13: HI Actual Vote Mode Shift for Individual Voters, 2020-2022

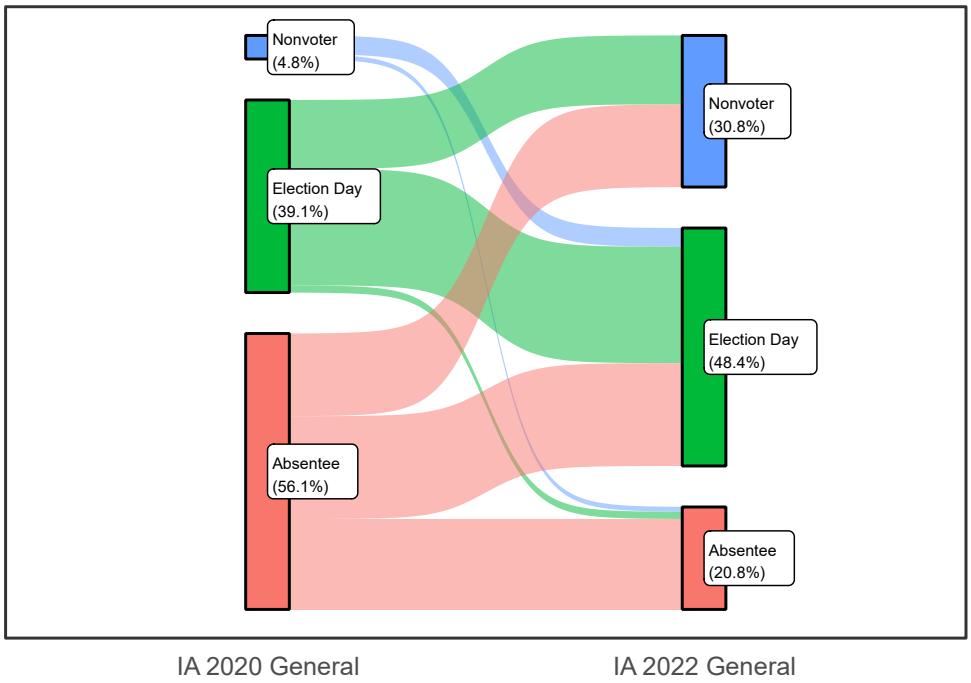


Figure 14: IA Actual Vote Mode Shift for Individual Voters, 2020-2022

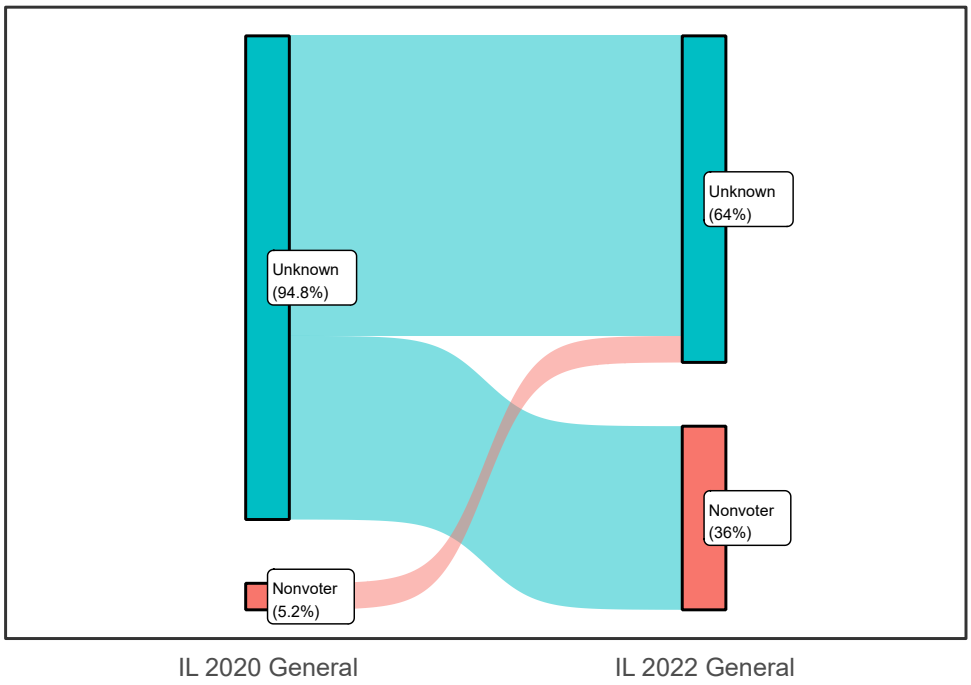


Figure 15: IL Actual Vote Mode Shift for Individual Voters, 2020-2022

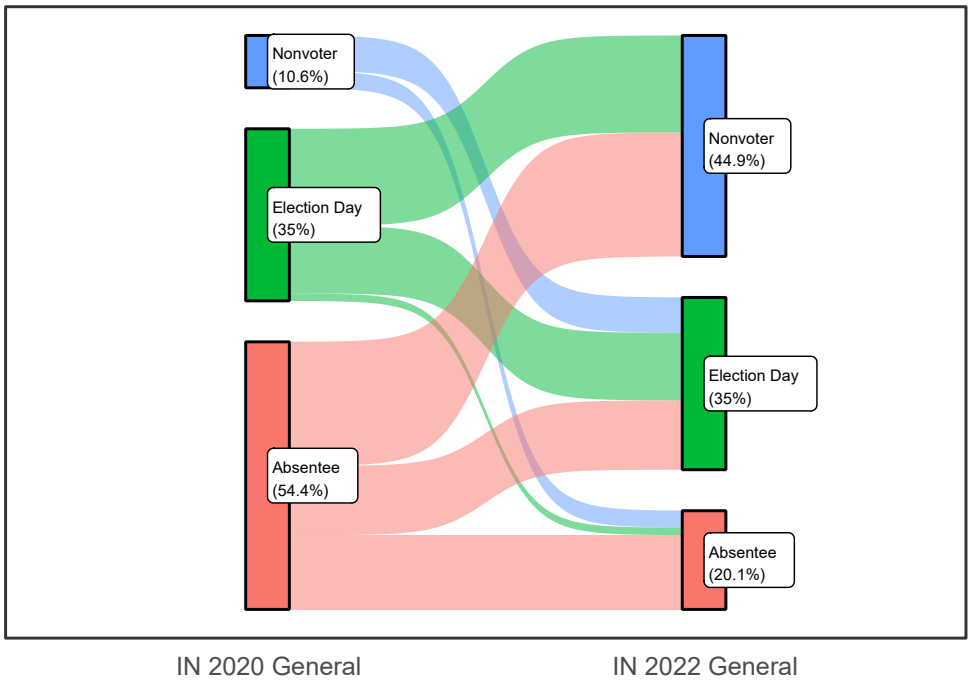


Figure 16: IN Actual Vote Mode Shift for Individual Voters, 2020-2022

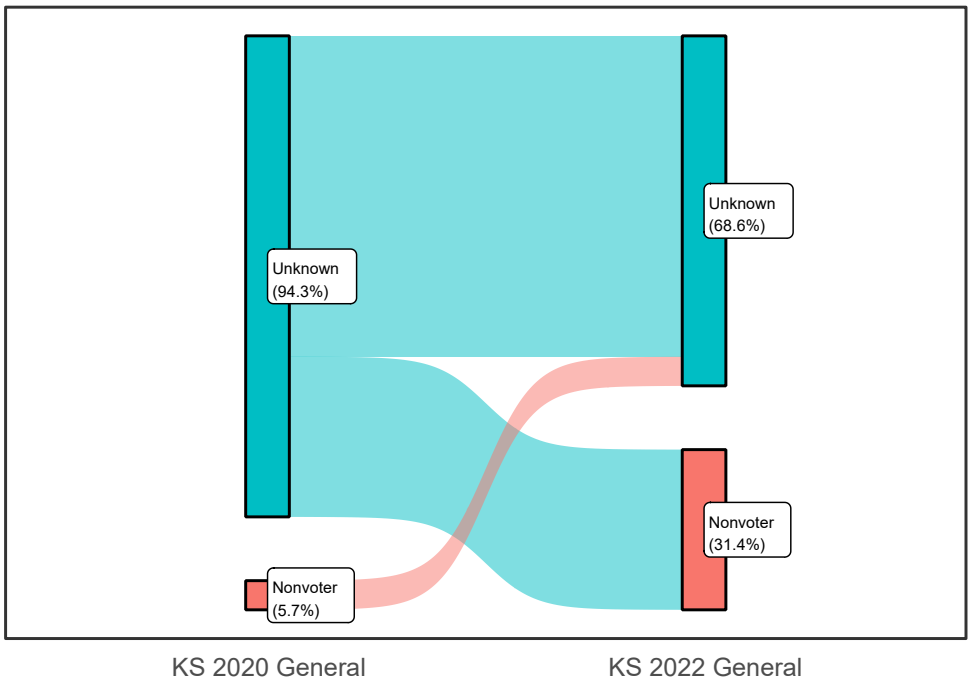


Figure 17: KS Actual Vote Mode Shift for Individual Voters, 2020-2022

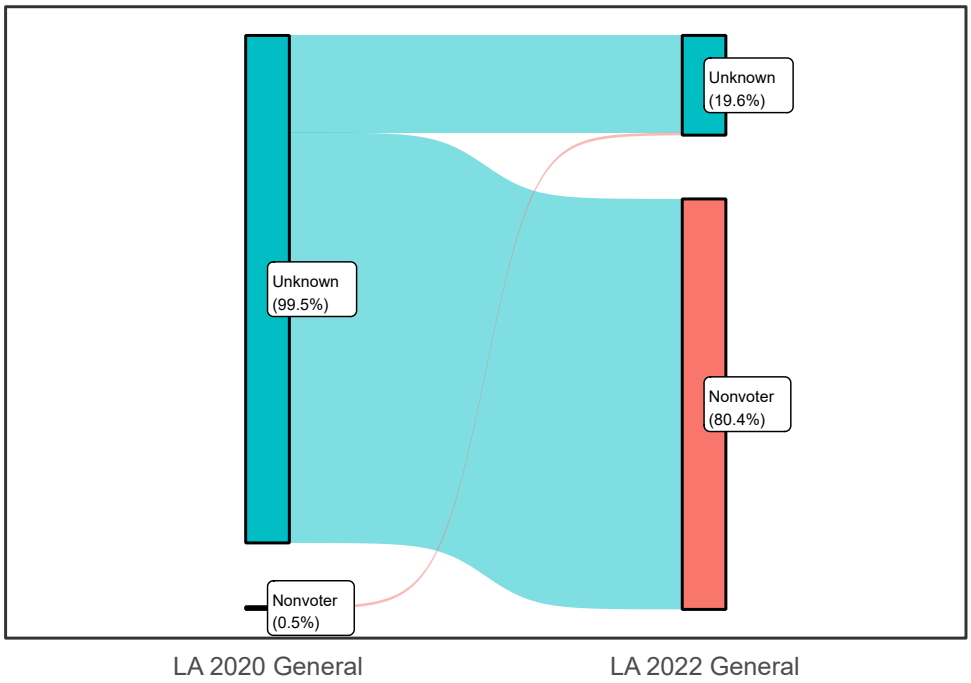


Figure 18: LA Actual Vote Mode Shift for Individual Voters, 2020-2022

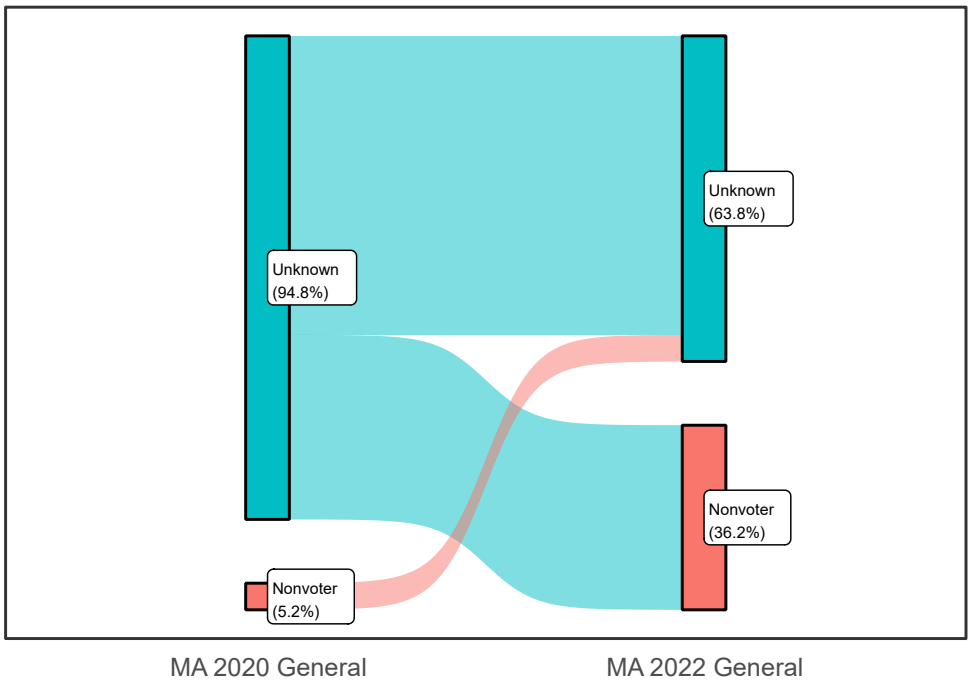


Figure 19: MA Actual Vote Mode Shift for Individual Voters, 2020-2022

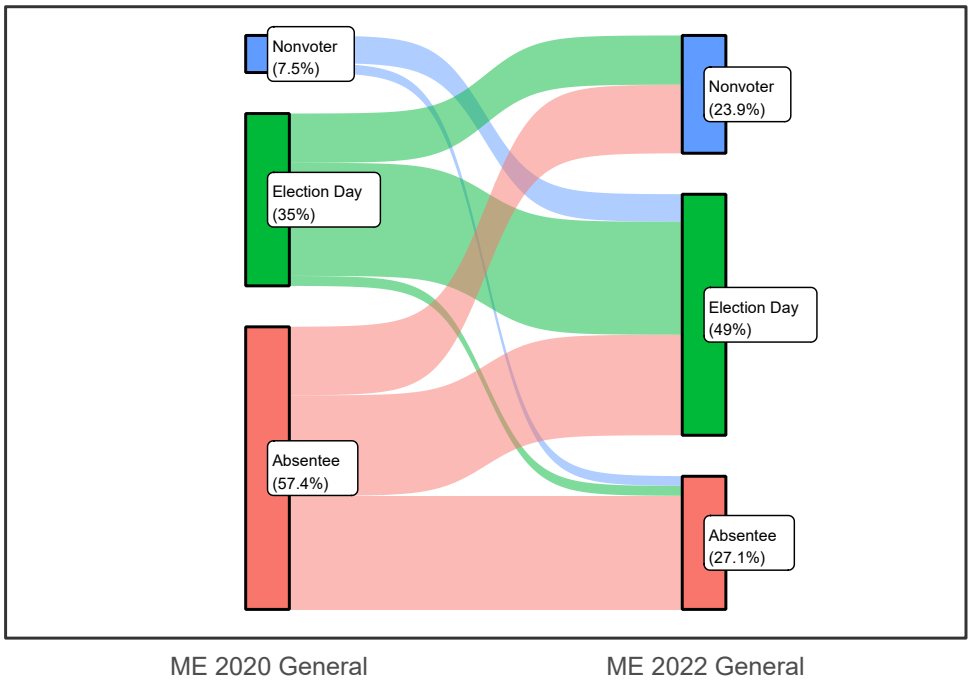


Figure 20: ME Actual Vote Mode Shift for Individual Voters, 2020-2022

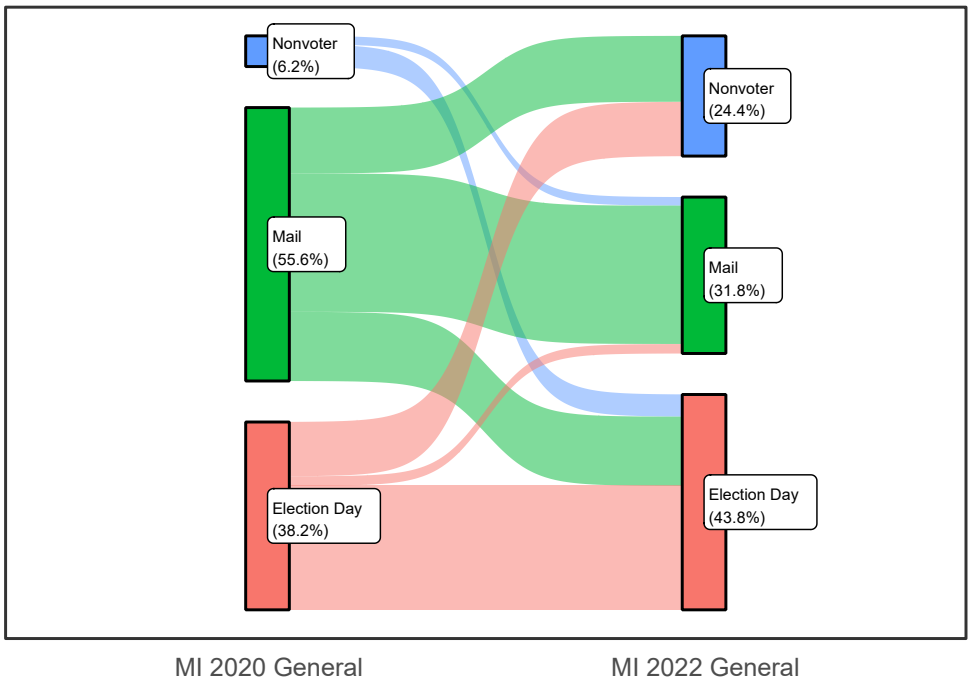


Figure 21: MI Actual Vote Mode Shift for Individual Voters, 2020-2022

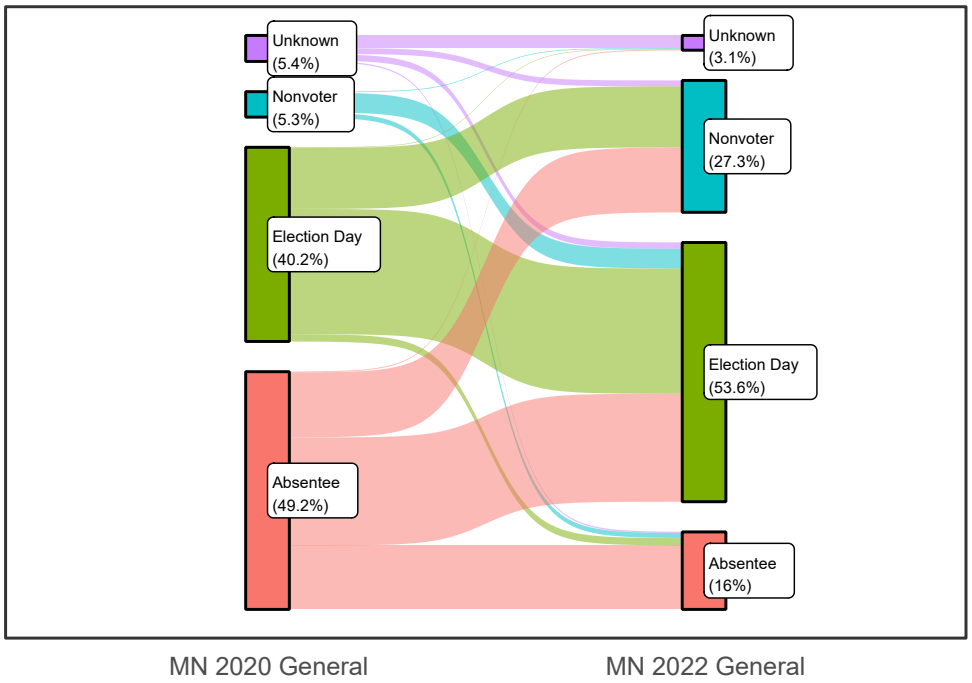


Figure 22: MN Actual Vote Mode Shift for Individual Voters, 2020-2022

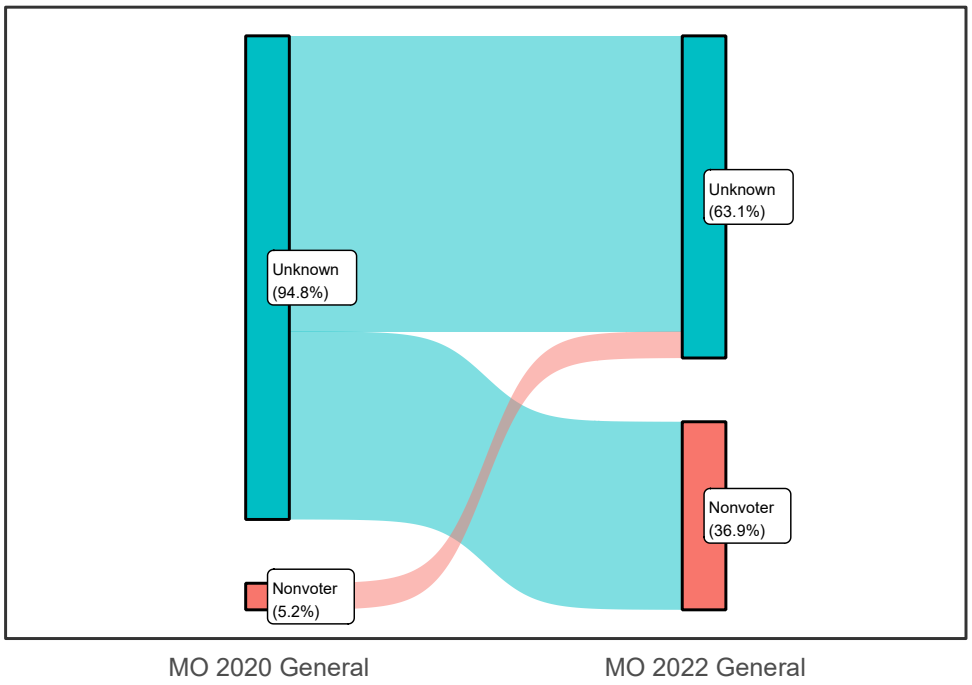


Figure 23: MO Actual Vote Mode Shift for Individual Voters, 2020-2022

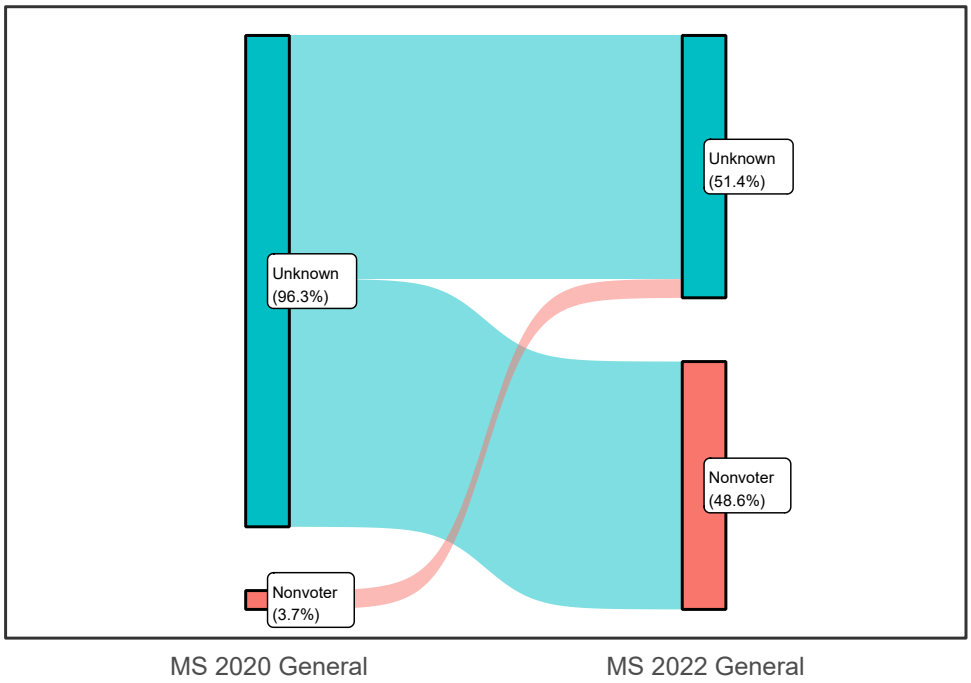


Figure 24: MS Actual Vote Mode Shift for Individual Voters, 2020-2022

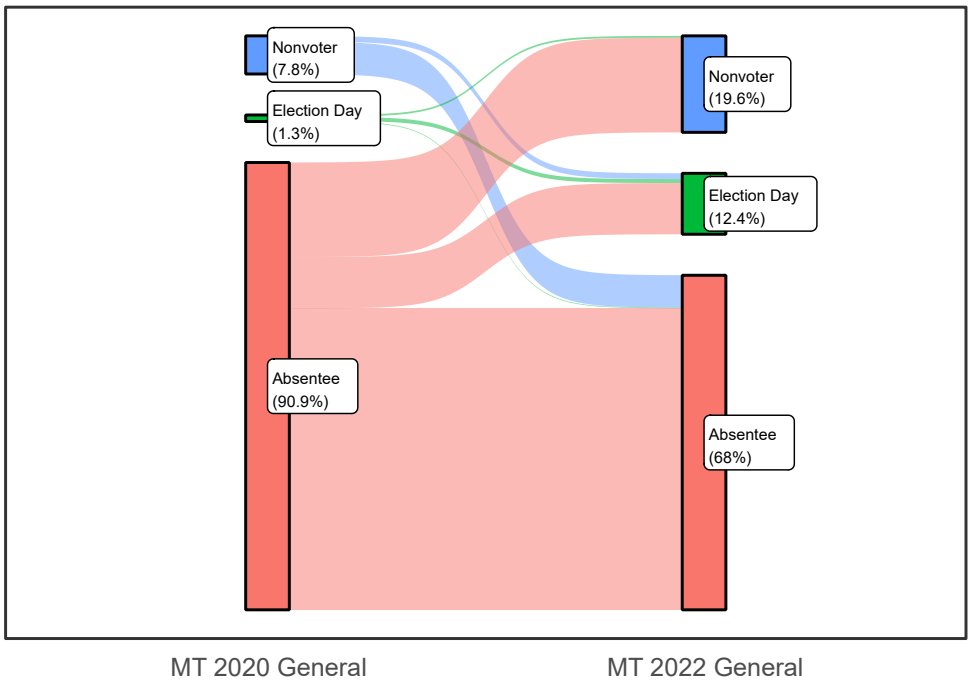


Figure 25: MT Actual Vote Mode Shift for Individual Voters, 2020-2022

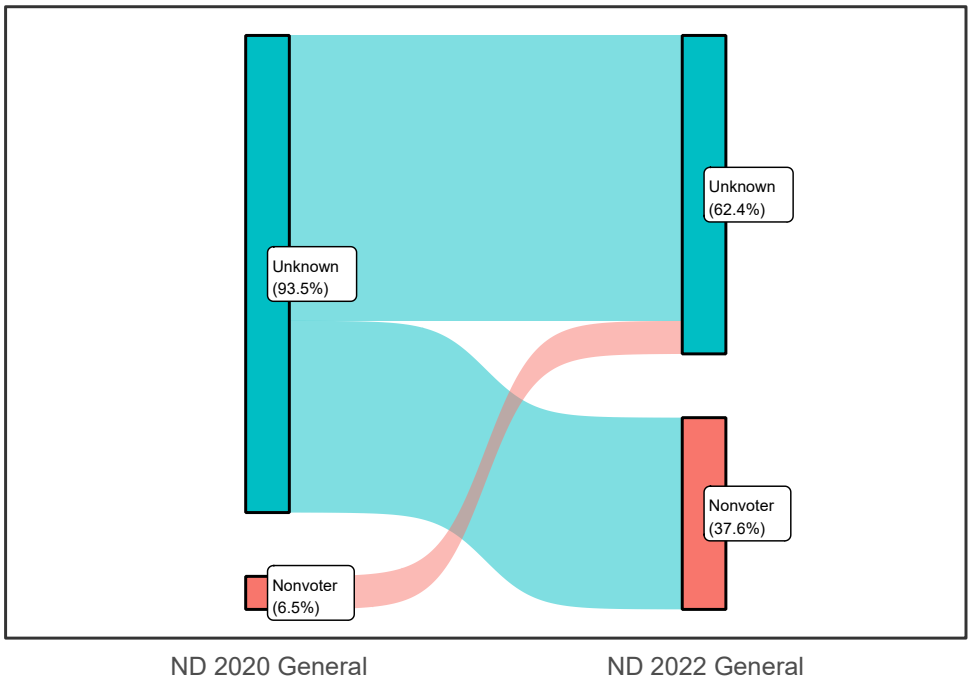


Figure 26: ND Actual Vote Mode Shift for Individual Voters, 2020-2022

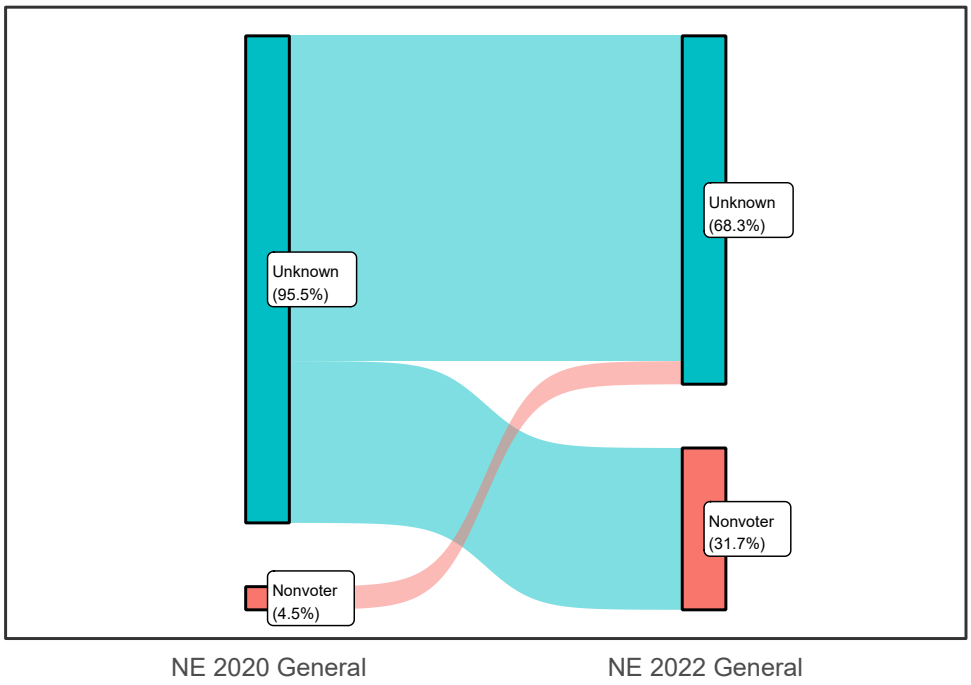


Figure 27: NE Actual Vote Mode Shift for Individual Voters, 2020-2022

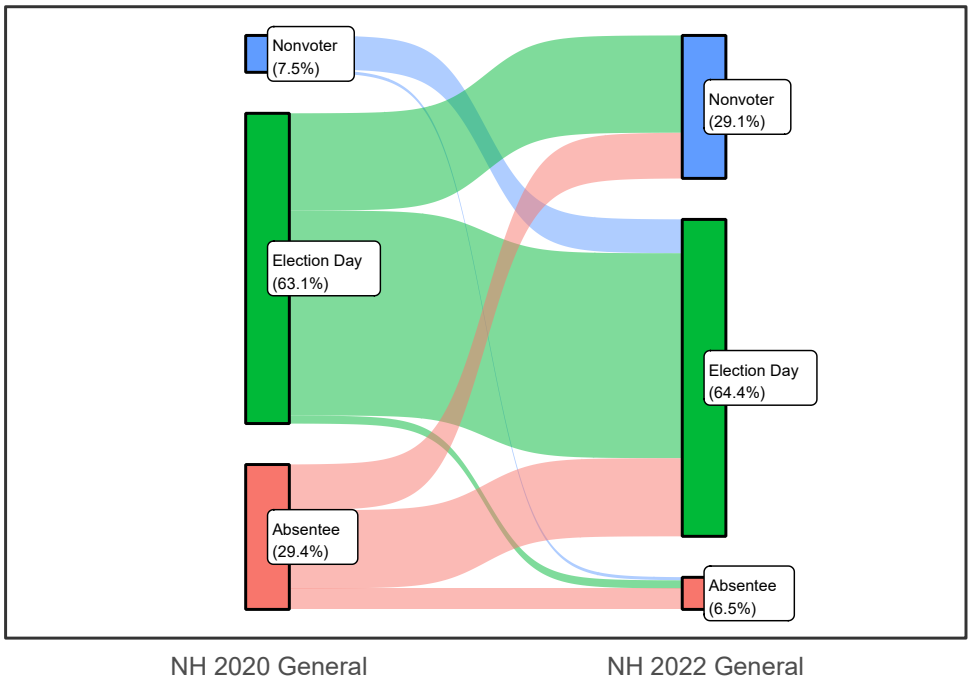


Figure 28: NH Actual Vote Mode Shift for Individual Voters, 2020-2022

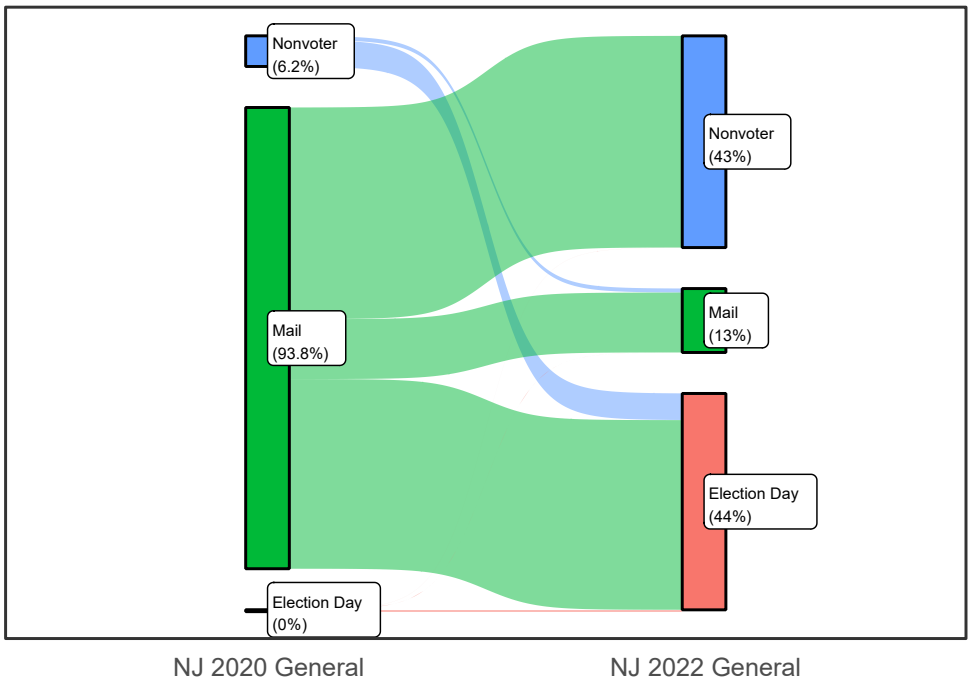


Figure 29: NJ Actual Vote Mode Shift for Individual Voters, 2020-2022

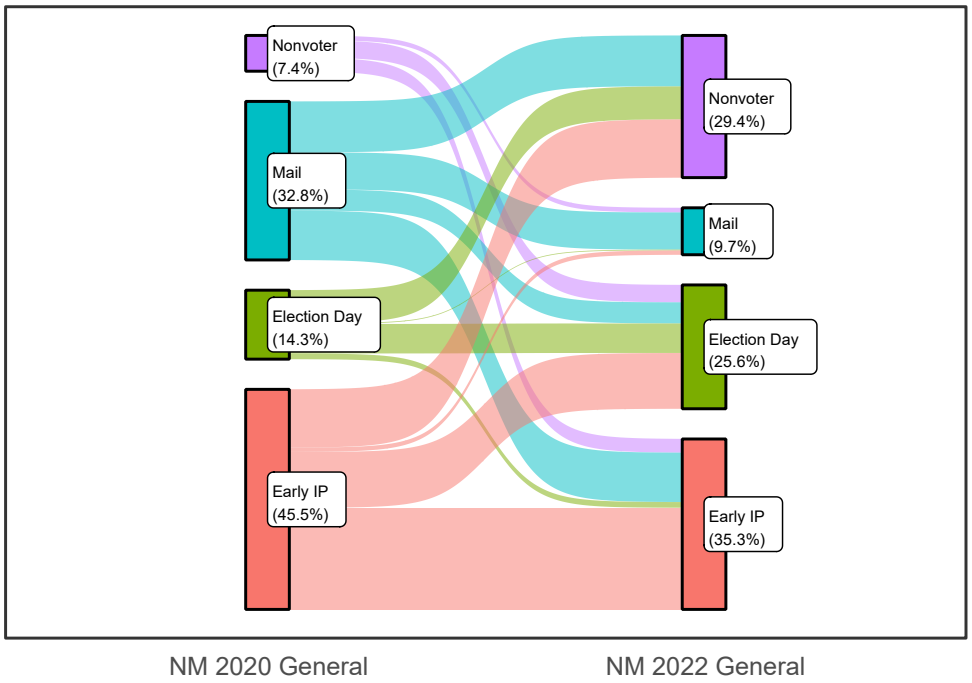


Figure 30: NM Actual Vote Mode Shift for Individual Voters, 2020-2022

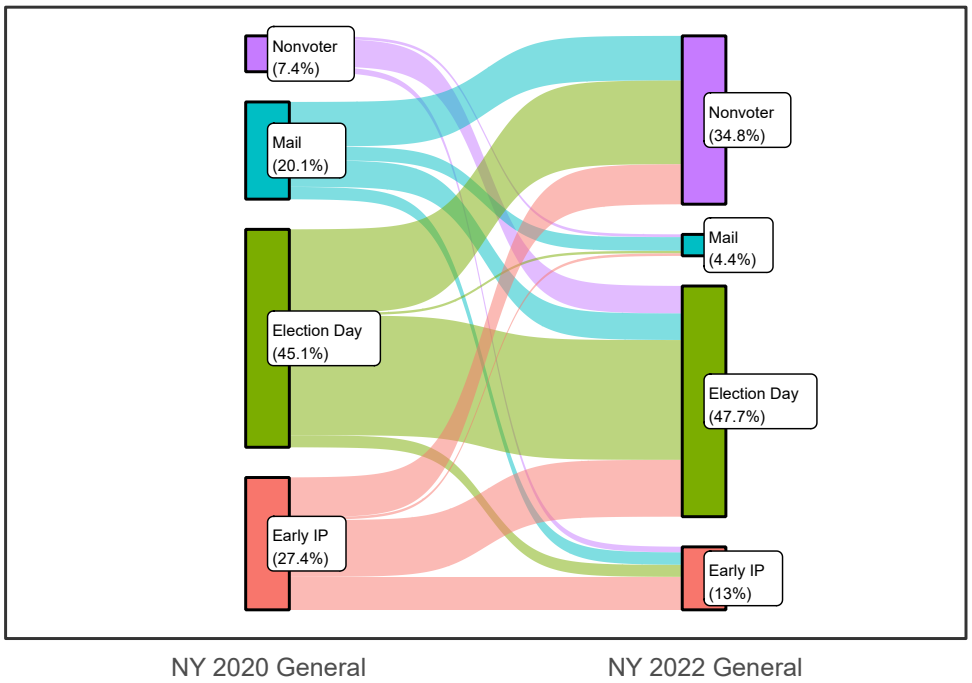


Figure 31: NY Actual Vote Mode Shift for Individual Voters, 2020-2022

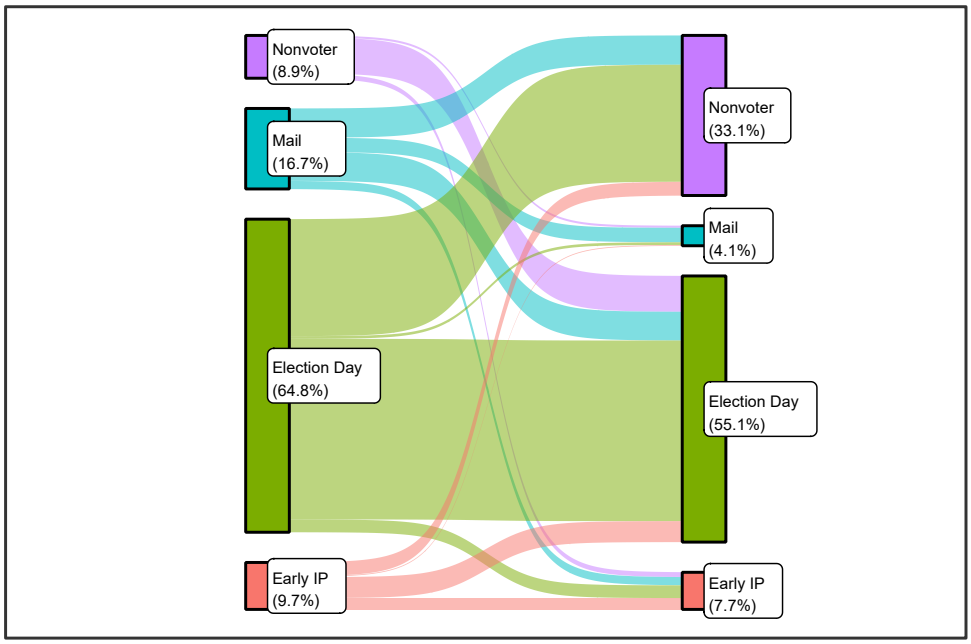


Figure 32: OK Actual Vote Mode Shift for Individual Voters, 2020-2022

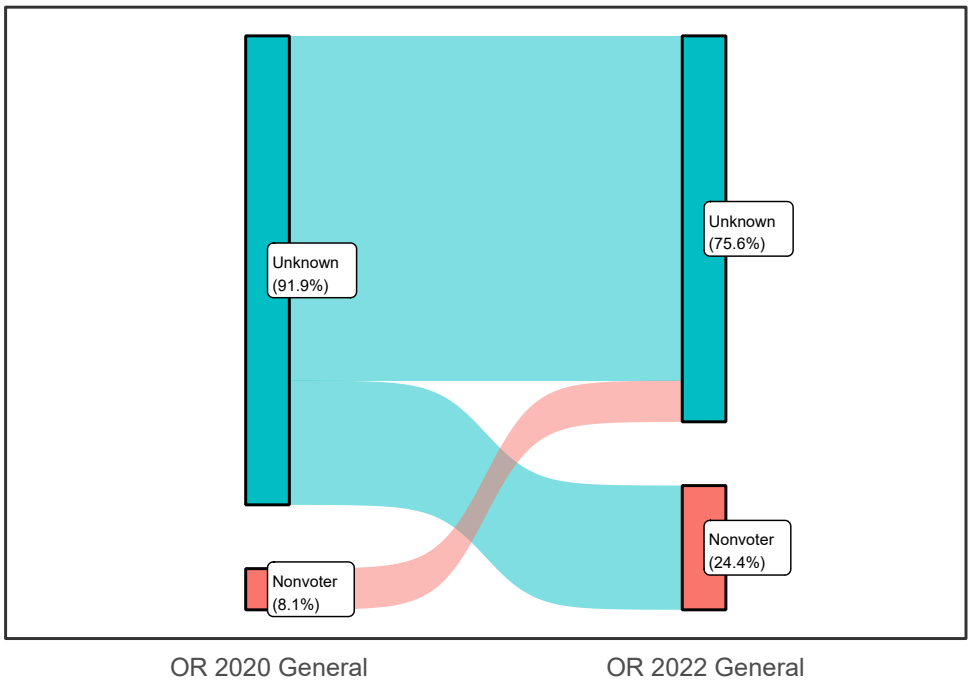


Figure 33: OR Actual Vote Mode Shift for Individual Voters, 2020-2022

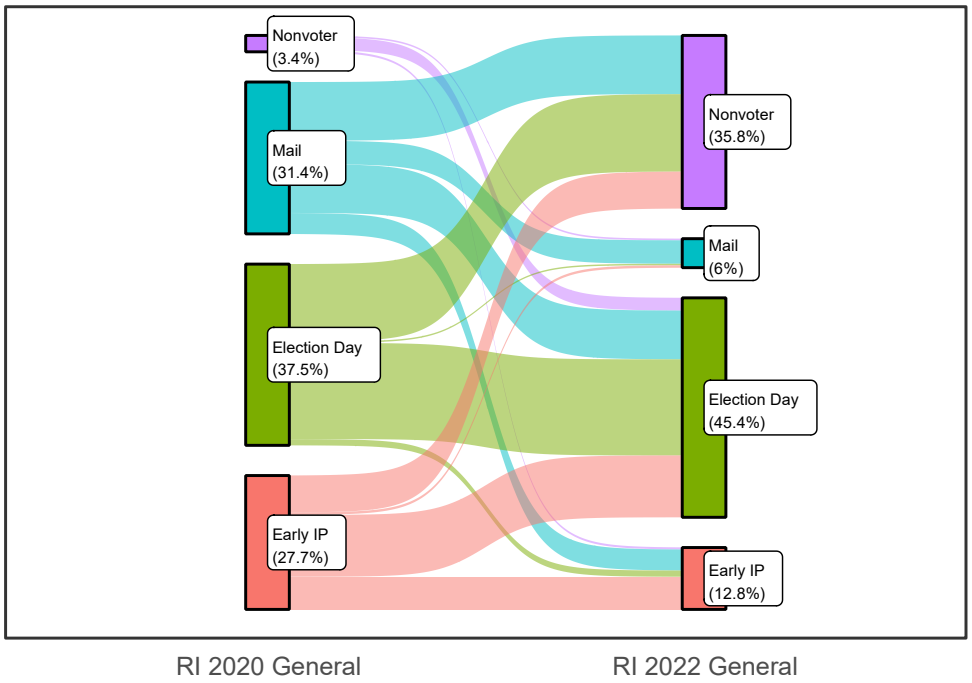


Figure 34: RI Actual Vote Mode Shift for Individual Voters, 2020-2022

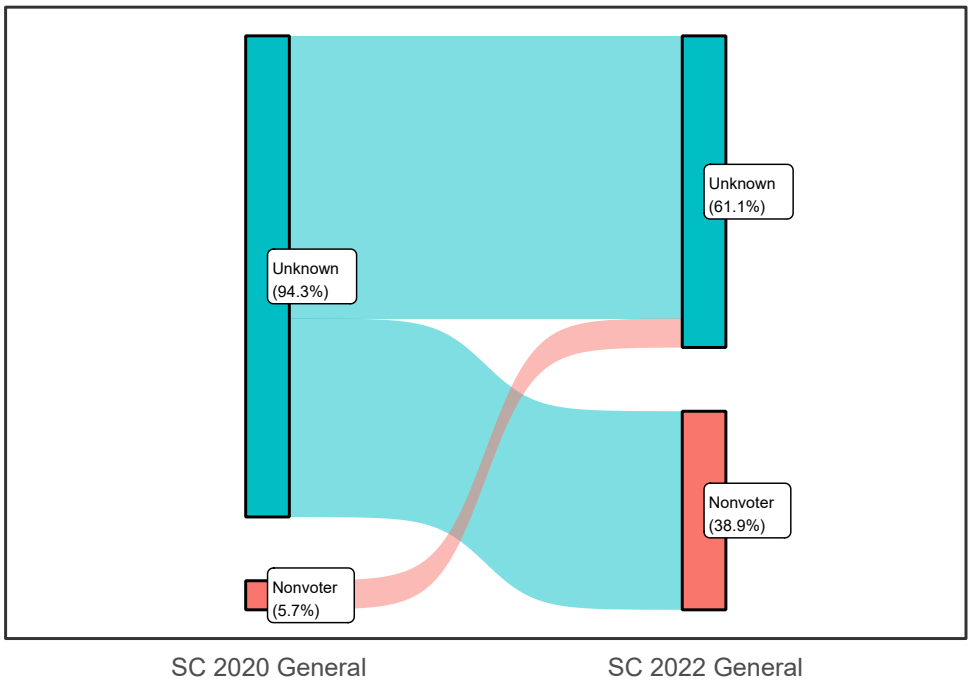


Figure 35: SC Actual Vote Mode Shift for Individual Voters, 2020-2022

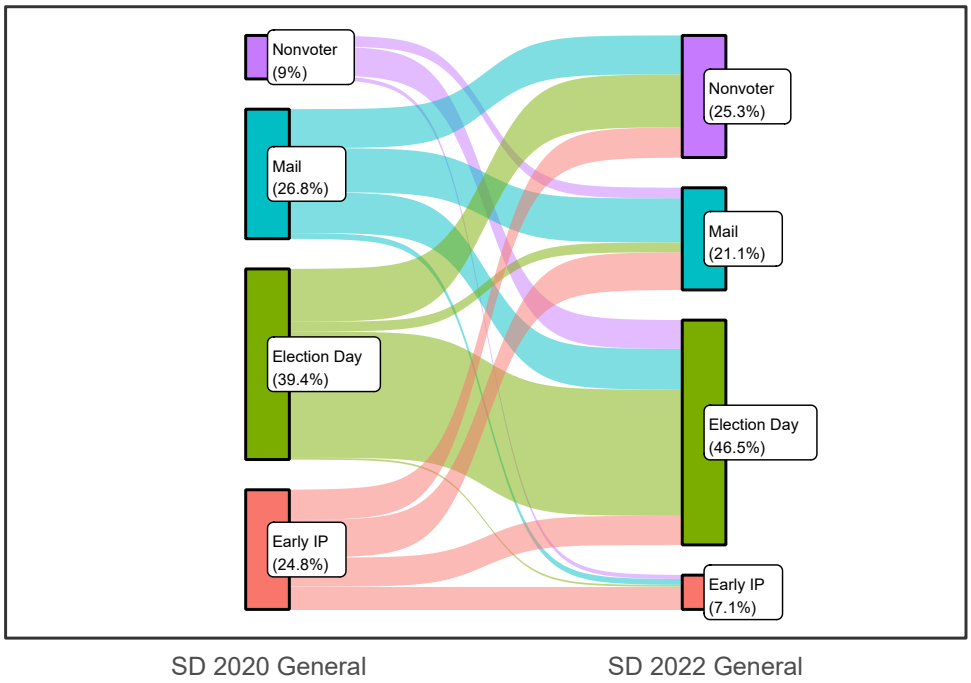


Figure 36: SD Actual Vote Mode Shift for Individual Voters, 2020-2022

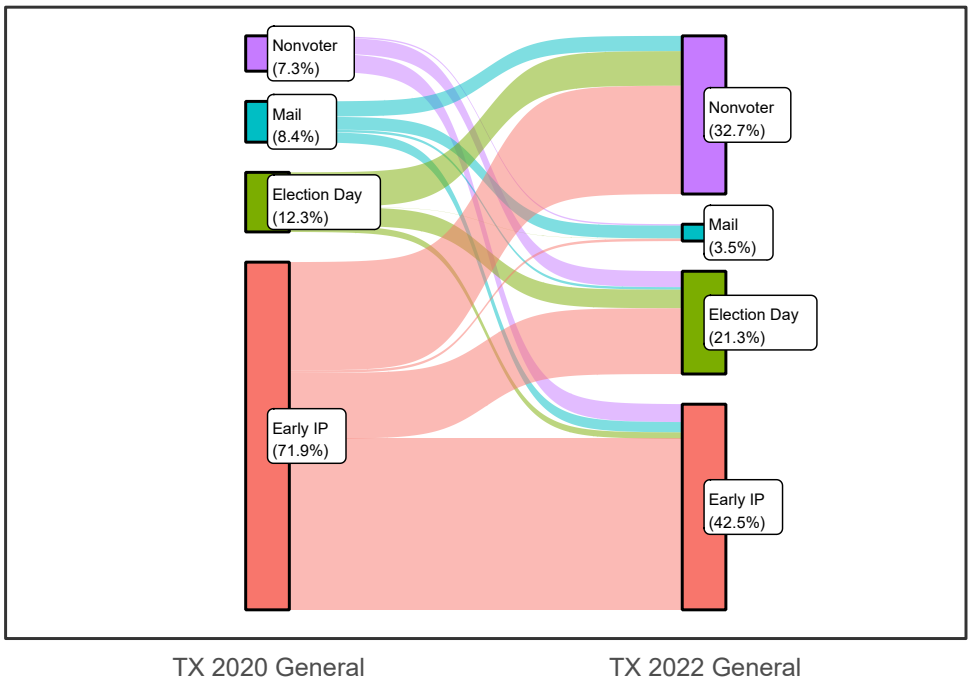


Figure 37: TX Actual Vote Mode Shift for Individual Voters, 2020-2022

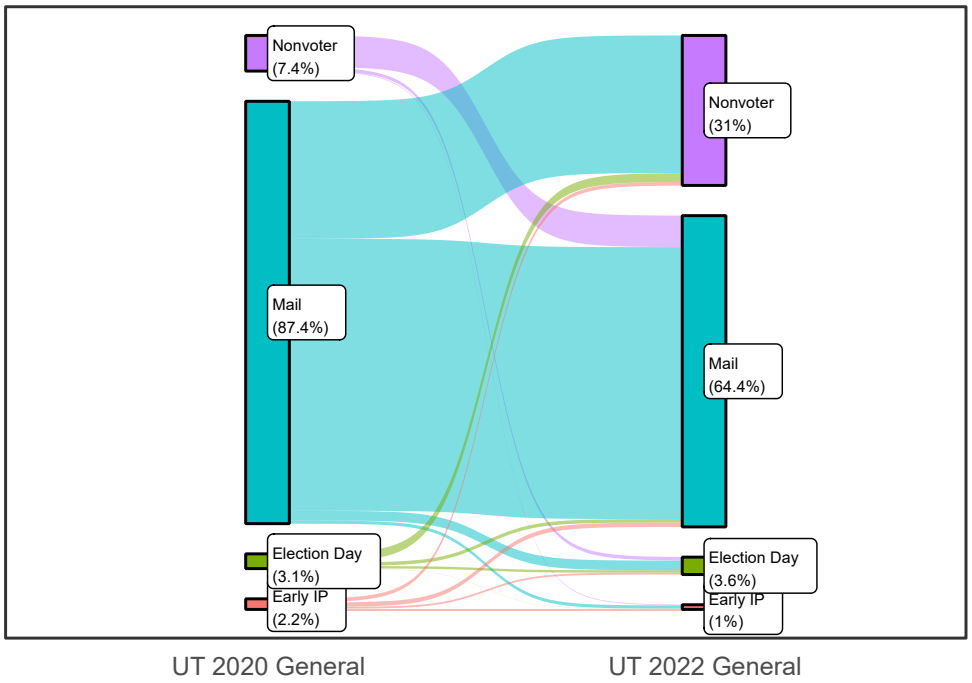


Figure 38: UT Actual Vote Mode Shift for Individual Voters, 2020-2022

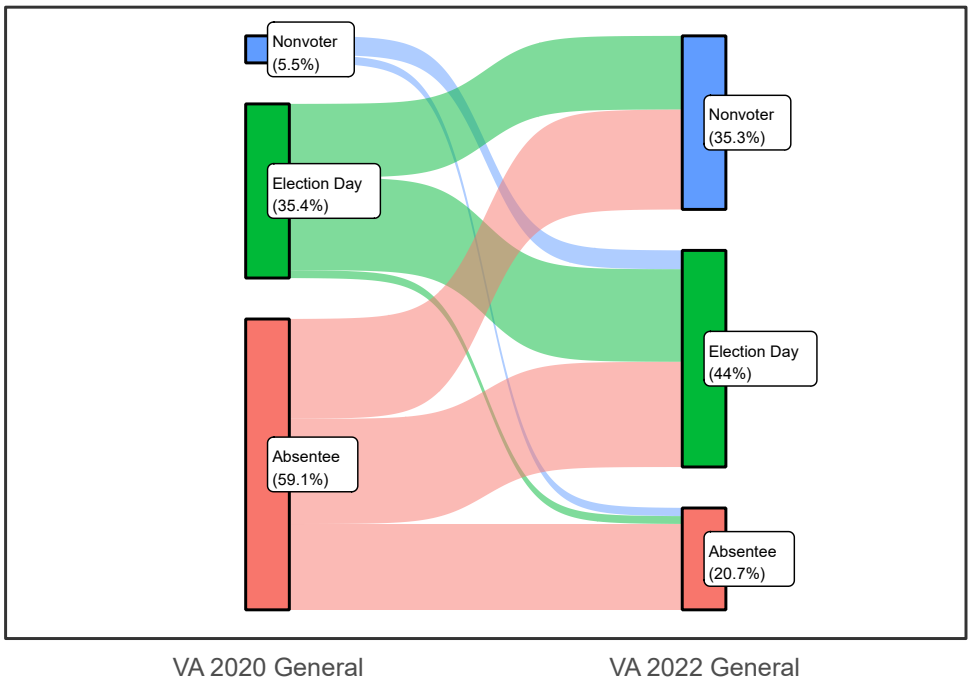


Figure 39: VA Actual Vote Mode Shift for Individual Voters, 2020-2022

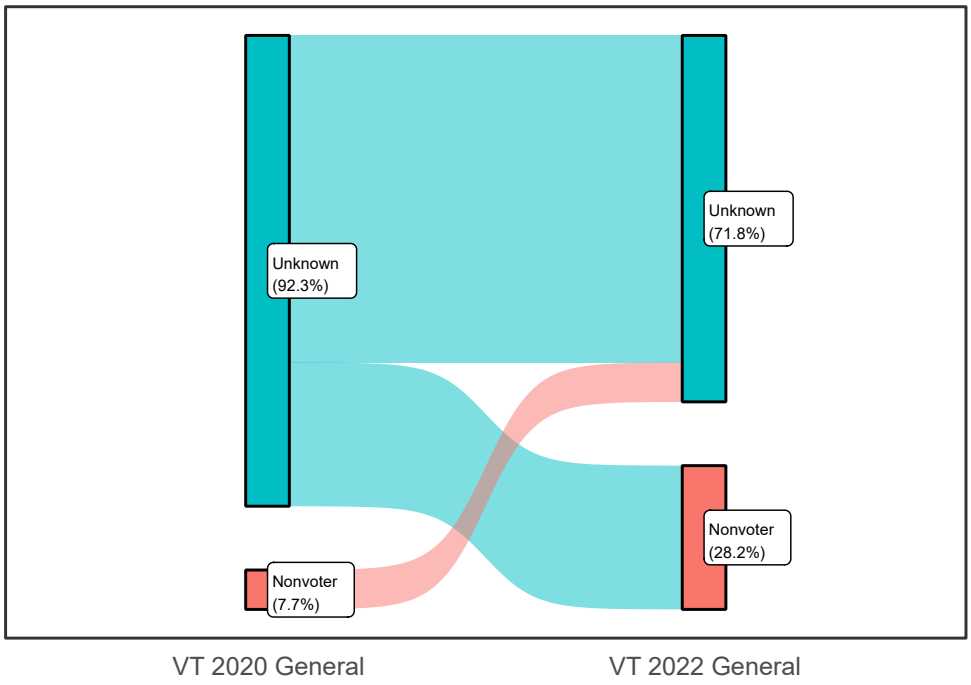


Figure 40: VT Actual Vote Mode Shift for Individual Voters, 2020-2022

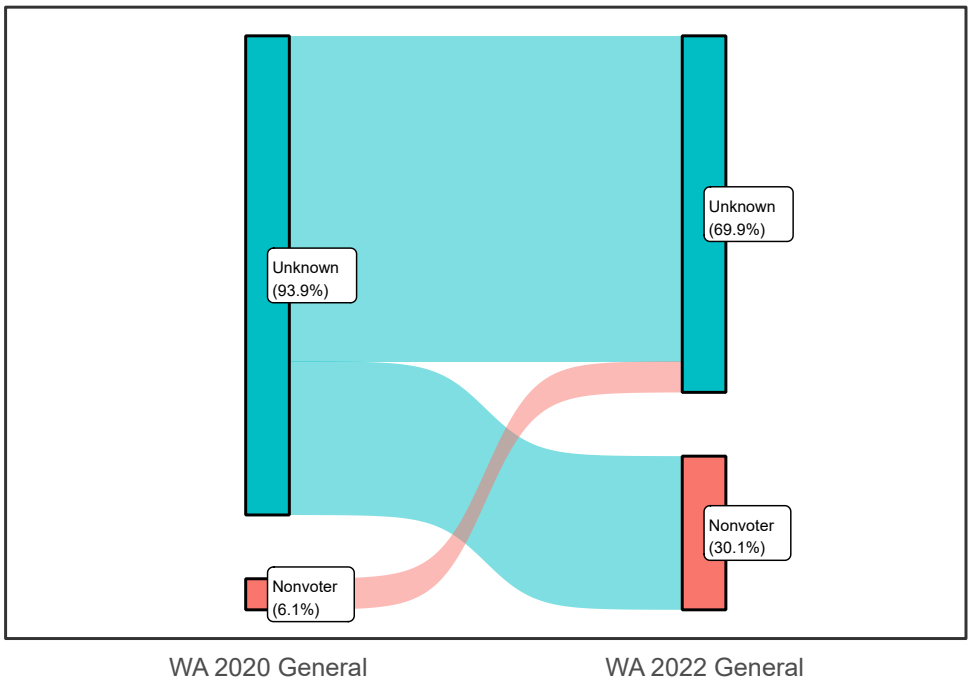


Figure 41: WA Actual Vote Mode Shift for Individual Voters, 2020-2022

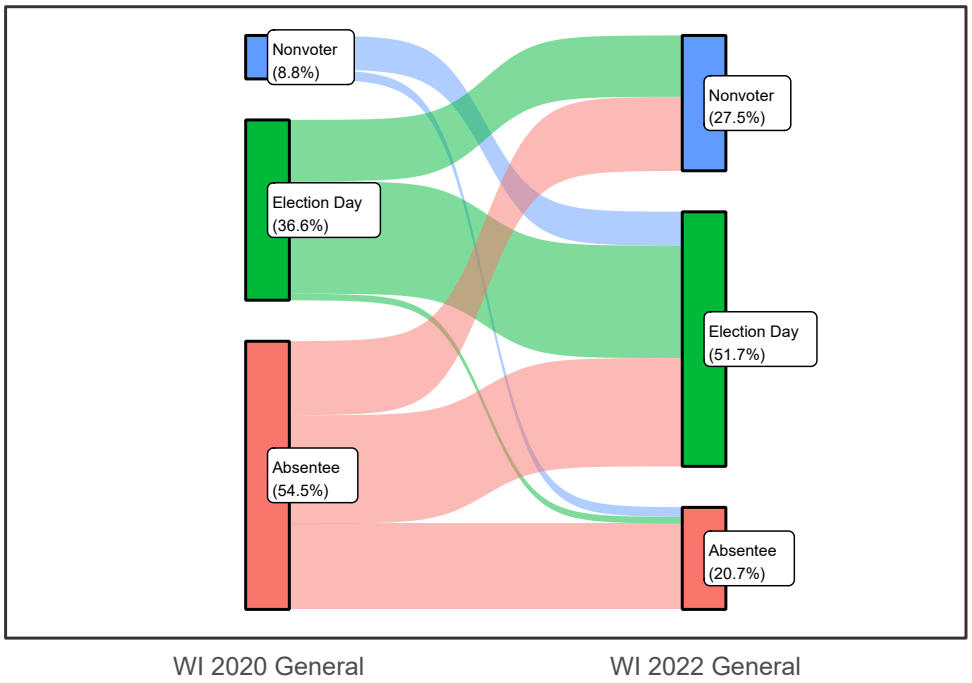


Figure 42: WI Actual Vote Mode Shift for Individual Voters, 2020-2022

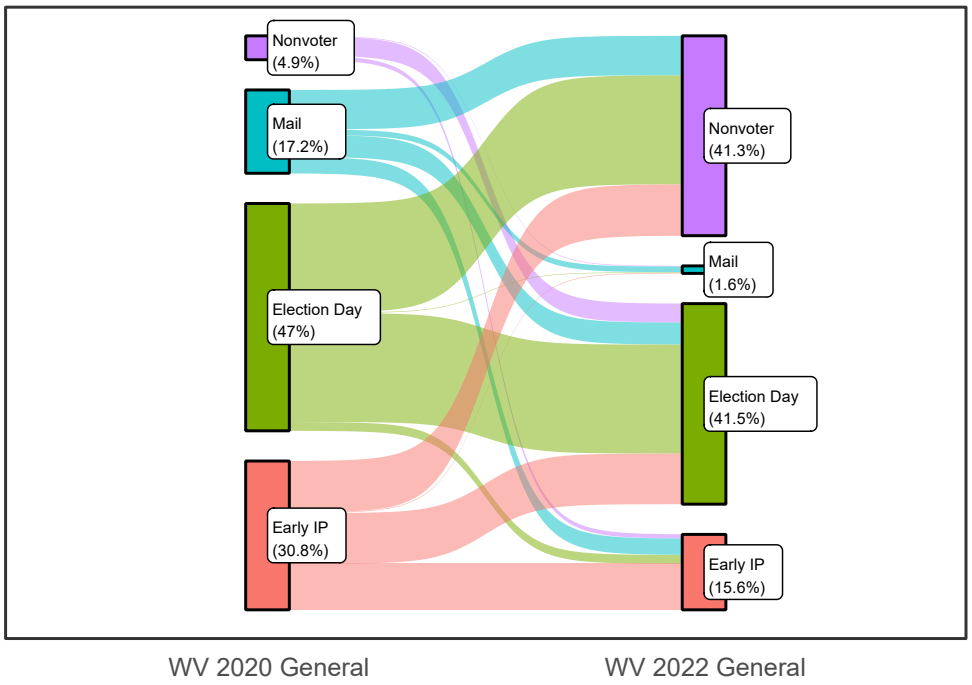


Figure 43: WV Actual Vote Mode Shift for Individual Voters, 2020-2022