The Institute
for Quantitative Social Science
at Harvard University

The Quality of Voter Registration Records: A State-by-State Analysis

Stephen Ansolabehere and Eitan Hersh

Department of Government
Harvard University

July 14, 2010

This paper is published as a working paper by the Institute for Quantitative Social Science at Harvard University and by the Caltech/MIT Voting Technology Project. Please direct correspondence to Stephen Ansolabehere at sda@gov.harvard.edu and to Eitan Hersh at hersh@fas.harvard.edu. The authors thank Bob Blaemire, Laura Quinn, Taylor Terry, and the staff of Catalist for their helpful guidance and cooperation.

Abstract

Voter registration systems in the United States have long been viewed as the area of election administration most in need of improvement. Problems in this system create barriers to voting for many Americans, but they also make it difficult for administrators to communicate with voters, identify voters at the polls, and audit elections after the fact. Improving the system first requires knowing the landscape and magnitude of potential problems. This report presents the first comprehensive, nation-wide analysis of the quality of information stored on voter registration lists. We offer a snapshot of the lists as of 2010 using data provided to us by the firm Catalist, one of the nation's leading vendors of voter registration data to political campaigns, organizations, and researchers. Catalist compiles all state and county election lists, standardizes those lists, and checks the accuracy of the information against other sources, such as National Change of Address registry and the Postal Service list of valid addresses. We examine indicators of the accuracy of eleven different pieces of information in the Catalist voter files. These indicators address four distinct uses of registration information: (1) assigning voters to precincts and communicating with voters, (2) validating people at the polls, (3) keeping lists current, and (4) auditing election results.

On the whole the picture that emerges is encouraging. (1) Approximately 4 percent of addresses on voter files are incomplete or invalid. (2) Identifying information such as birthdates are generally well collected, but several states do not have such identifying data and there are irregularities in six states. (3) In the typical state approximately 4 percent of records are obsolete (usually because the person has moved) and the proportion of deceased people on lists is small, less than one percent in most states. These rates compare favorably with measures such as obsolete addresses in the Census enumeration efforts. (4) Discrepancies exist between voters recorded as voting and ballots counted in most states, with about 2 percent of voters wrongly counted as having either voted or abstained. Somewhat surprisingly, there is little correlation among these indicators across states, though a few of states routinely score well and a few routinely perform poorly. This suggests to us that the most helpful information to states and counties is not a single performance indicator and rank of jurisdictions but more detailed information about the absolute level of accuracy of information and the comparison with other jurisdictions.

Summary of Key Results

- Of the 185,445,103 listed registration records in the United States, 16,130,325 are estimated to be invalid.
- Aside from invalid records, in the typical state 1 in 65 records is duplicative, meaning that the same registrant is listed multiple times.
- 1 in 25 records contains a mailing address that is likely to be undeliverable because of a typo, a street that no longer exists, or poor penmanship on registration applications.
- In the typical state, 1 in 40 counted votes in the 2008 general election cannot be matched to a registrant listed as having voted.
- 1 in 100 listed registrants is likely to be deceased.
- 1 in 7 records does not have a listed birthdate, and for many voters who do have a listed birthdate, the date entered is inaccurate.
- 1 in 25 registration records is estimated to be deadwood, because of registrants who have not voted in a very long time, have moved elsewhere and re-registered, or are thought to be deceased.
- 1 in 60 registrants do not have a date of registration associated with their record, and and implausibly large number of registrants who do have a registration date (1 in 50) are listed as registering on January 1st.
- Measures of registration quality are uncorrelated; states that perform well on some measures perform poorly on others. States cannot be reasonably ranked on overall performance quality.

Introduction

Voter registration is the backbone of election administration in the United States. Registration lists are used to establish eligibility to vote, to determine the offices for which one can vote, to communicate to citizens when elections occur and where and how to vote, to validate people at the polls, and to audit elections after the fact. In many ways U. S. voter registration is a remarkable system. A highly decentralized set of authorities consisting of over 5,000 municipal and county election offices assemble and manage their own local lists, and these separate lists are aggregated into databases of all registered voters in each state. Although no governmental entity maintains a national database, the political parties and political data management firms assemble the state lists to create a comprehensive database of all 180 million eligible voters. This massive data collection enables parties and interest groups to identify and mobilize potential supporters.

Election administrators and the public as a whole place a premium on accuracy of the lists. Poorly maintained lists can make it difficult for administrators to communicate with voters or to run the election at the polls. Errors in the lists used by the local election offices can prevent some legitimate voters from participating and may be abused by those seeking to perpetrate voter fraud. Election administrators devote considerable effort to continual management of the voter lists. This is a difficult task, especially in less well staffed offices, and errors do occur. Any effort to improve the quality of lists can be greatly informed by data about the accuracy of the voter files. How common are different sorts of problems; do errors follow systematic (and possibly correctable) patterns; and what are the consequences of different sorts of errors?

To date, understanding the quality of the registration system has relied on surveys of individuals conducted by the U. S. Census and other organizations, and on intensive studies of individual counties or states.[1] As far back as the 1970s, such research has documented that registration systems create barriers to participation that affect several million Americans each election,[2] but the surveys have not had sufficient sample sizes and statistical power to determine the origins and magnitude of specific problems, or to compare jurisdictions. Examination of the system failures during the 2000 election identified voter registration as the single largest source of problems for voters in that election. Such revelations prompted the federal government to pass the Help America Vote Act in 2002, and they continue to motivate calls to improve the voter registration system today, including proposals to adopt universal registration and to force local governments to improve by ranking their performance.[3] But even the post-2000 assessments (and the reform efforts that have followed in their wake) have been unable to offer a comprehensive assessment of the quality of voter registration lists in

[1]See, for example, Raymond E. Wolfinger and Steven J. Rosenstone, *Who Votes?* New Haven: Yale University Press, 1980;; The Caltech/MIT Voting Technology Project; Allan J. Lichtman and Samuel Issacharoff, "Black/White Registration Disparities in Mississippi," *Review of Litigation* 13 (1993): 1-29.

[2]See Wolfinger and Rosenstone, 1980.

[3]For example, see Heather Gerken, *The Democracy Index*, Princeton: Princeton University Press, 2009; R. Michael Alvarez, "Measuring Election Performance," VTP Working Paper #94, Caltech/MIT Voting Technology Project, 2009; Pew Center on the States, "Data for Democracy," December 2008.

the United States.

This report presents the first nation-wide view of the accuracy of information on the voter lists. We examine the voter registration files of all the states to provide a comprehensive assessment of the validity of nearly a dozen specific pieces of information on the voter files, such as addresses, birthdates and vote histories. The novelty of this analysis owes to the fact that these data themselves are quite new. In just the last few years, political data management firms, working with parties, campaigns, and other organizations, have accumulated all state and local voter files and have compiled the first truly national voter lists for the United States. In addition to the availability of the first national registration file, tremendous advances in computing memory and computing power over the past decade make it possible to access and analyze all 180 million records on the voter files and to compare them with other comprehensive data about the American public. The data for this assessment comes through a contract with one particular firm, Catalist, LLC. Our link to this organization came about through research we conducted for the Pew Center on the States. Although Catalist vends data to Democratic party campaign organizations, labor unions, and other groups, we have no partisan or ideological agenda in pursuing this research, nor do we have any consulting relationship with this firm.

With so many different pieces of information that may be validated, it is easy to get lost in the details. It is tempting to construct a single measure of list quality, as is advocated by Professor Gerken of Yale Law School. However, voter registration lists do not have a single purpose and not all information on the lists is immediately relevant to each purpose. Furthermore, our research indicates that the registration problems states face are not closely related to one another, which suggests that a summary measure may obscure more than it would reveal. We thus organize this assessment around what we see as four distinct functions of voter files.

The first function is to map voters to precincts. To identify a voter's correct precinct, administrators need to know the address of the voter. When precincts are established, a list of all voters eligible to vote in that precinct (and to vote for the unique set of offices corresponding to that precinct) is generated. This list is used by local election offices, campaigns, and other organizations to communicate with registered voters about where, when, and how to vote. This function requires complete, mailable addresses. An address with just a street name or an address that does not correspond to a known geographic location makes it impossible to place voters in precincts.

The second function of a registration list is to authenticate voters at the polls. When a person goes to vote, poll workers verify that the person is indeed allowed to vote at that location. In every state (but North Dakota), the individual must say where he or she lives, so that the poll worker can check that the person is in the right location. Half of the states require some form of identification when voting. Names, addresses, and birthdates are used to validate a registrant's identity. A voter whose identifying data does not match registration records may encounter problems at the polls.

The third function of registration lists is to audit election results. One part of election audits involves comparison of the lists and the vote tallies: does the total number of people

recorded as having voted equal the number of ballots cast?

The fourth function is to prevent fraud. Duplicate and obsolete records allow for the possibility that people vote multiple times in several places. To maintain clean lists, states must identify people who are registered in multiple locations in a state as well as those who have moved out of the state or have passed away. States may also purge voters following the procedures laid out in the National Voter Registration Act and in state laws. Under these rules, registrants who fail to vote for a series of elections can be sequestered and eventually dropped. In order to purge voters on the basis of past voting participation, administrators must maintain accurate vote history data as well as an accurate record of when voters entered the registration system.

The bulk of this report is organized around these four functions and provides a state-by-state analysis of election administration based on the accuracy of nearly a dozen pieces of information on the lists. It is the first analysis of its kind. As with any new measurement tool, future refinements and improvements will be necessary. This analysis covers facets of the election administration related to voter registration records. Other important aspects of election administration, such as the counting of ballots, are left to other research.

We begin by describing the source of data that provides the basis for our evaluation. We then describe the eleven measures we have developed to assess registration records. We review national-level results on these measures and then turn to state-by-state comparisons.

The Data

Data for this analysis come from an emerging commercial business. A government-managed national voter registration system does not yet exist, but privately managed national lists have emerged to fill this void. Political campaigns, such as those conducted by the national political parties and interest groups, use public registration records as the basis for their voter contact programs. Precisely because registration records are a burden to collect and contain inaccuracies, campaigns purchase lists from professional data vendors. The vendors do the painstaking work of collecting the data, flagging address errors and typos, and identifying duplicative records, deceased voters, and movers. There are two by-products of this effort. First, the commercial vendors have, in essence, created national voter registration lists. Second, they have conducted audits of the public lists available from local and state offices. Election data firms do this as a means of lowering costs of campaigning, for example by getting rid of duplicate mailings or mailings to people no longer at an address. Because their customers want clean lists so that they can determine which voters to contact and which records are obsolete, these vendors have developed extensive understanding of the quality of registration lists originating from the state and county election offices. In this analysis, we draw inferences about the quality of registration records by investigating how a state's voter registration list is transformed into a list in which obsolete records are removed and inaccuracies are identified. We examine the election data assembled by the firm and the firm's own indicators of problematic information on the lists.

Catalist, LLC, a Washington based firm, is one of the premier political data vendors in the

business. From voter registration records in every state, Catalist's team builds a nationwide voter file that is used by the Democratic Party and by interest groups on the Democratic side of the political spectrum. During the winter and spring of 2010, we conducted a detailed investigation into Catalist's data cleansing process through an intensive study of their voter records and interviews with their senior staff.

Because of the nature of the business, Catalist, like most other political data vendors, serves only one political wing. However, after a careful study of their process, we are confident that the results described here are not affected by partisan bias. The reason for this confidence is simple: the vendor has neither cause nor much capacity to discriminate on the basis of partisanship when developing and cleaning the lists. Consider the basic method they use. Several times a year, Catalist purchases the publicly available voter registration files made available by each state or county election office. For every voter on a list Catalist makes a note of missing or duplicative information. Is the voter's address complete or is it missing a key piece of information such as a ZIP code or street address? Is the voter's birthdate absent from the file? Does the voter show up multiple times in the state because the registration record from a former residence was never deleted from the file?

Catalist then cross-references the registration lists with other public records, such as the National Change of Address (NCOA) database maintained by the Post Office and the Social Security Death Index. Movers and deceased voters are flagged. Catalist matches the registration files to commercial records from data aggregation firms that compile lists from retailers and direct marketing companies. This allows the firm to correct the records of individuals who may have a typo in their registration record or may have registered with a nickname rather than their legal name. For example, if a voter registered with the first name "Tom," but a collection of commercial records show that his real first name is "Thomas," Catalist will note the discrepancy.

Statistical data matching procedures are run on every voter file that Catalist processes. Catalist does not attempt to create more careful records for likely Democratic voters or likely Republican voters, nor does it attempt to conduct a more rigorous data cleansing of some states than others. Biased processing of this sort would result in less reliable data for Catalist's clients, who work to mobilize all kinds of citizens in every state, depending on the nature of independent issue campaigns. Thus, we are quite confident that analysis of Catalist's records is not biased by Catalist's political affiliations. In fact, because data vendors on both sides of the political spectrum use the same voter registration data and create lists with the same purposes in mind, we would expect an analysis of records of a Republican vendor's data to produce the same basic results. Of course, we would embrace an opportunity to study the records of a Republican vendor just as we have done with Catalist's records. We view partnerships with private data firms as a new and valuable approach to assessing public data, and we hope that it will be replicated in the future.

Data Quality Indicators

To assess the voter registration system in each of the four functions described above, we must estimate the accuracy of mailing address information, identifying information (such as birthdates), vote history records and records indicating when people entered the registration system, as well as the incidence of obsolete and duplicative records. We have developed eleven indicators to measure accuracy in each of these areas.

1. **Missing Address Information** The number of registrants missing a city, state, or ZIP code field in their mailing address. Missing this information will make it difficult for election administrators to send notifications about voting to a registrant. Additionally, a person's mailing address is usually the same as his or her registration address. Consequently, missing a city, state, or ZIP code could make it difficult to place a voter in a precinct.

2. **Undeliverable Addresses** An estimate of whether a mailing address is erroneous. Catalist runs every mailing address through a U.S. Post Office process called CASS, the Coding Accuracy Support System. The CASS process generates an estimate of how likely it is that an address is valid. Consider an example. Suppose a voter lives at 100 Main Street; however, on account of a typo or poor penmanship, the address is listed on the voter file as 109 Maln Street. Although this address is missing no values, the CASS process might find that Maln Street does not exist in this particular city, or that Maln Street does exist but that 109 is not likely to be a viable street number on Maln. In Catalist's data, the output of the CASS process is a five-category rating: mailable, probably deliverable, possibly deliverable, probably undeliverable, or undeliverable.

3. **Birthdate Coverage** The number of registered voters for whom Catalist does not have a record of date of birth from the voter files.

4. **Birthdate Distribution** The percentage of people listed as being born on each day of the year. Even when a birthdate is associated with a voter on the registration file, the birthdate might not be correct. One good signal of the accuracy of birthdate records is the distribution of voter birthdays throughout the year. Records should show that approximately the same number of registrants have a birthday on each day of the year. If a disproportionate number of voters have a listed birthdate on any one day (such as January 1st), this suggests inaccurately recorded data on the part of election administrators.

5. **Duplicative Records** The number of records Catalist removed from the registration file because they were duplicative within a state. Americans change residences very frequently, and most people who move stay within the same state, and usually within the same county. If citizens re-register when they move, then a state voter registration file will contain lots of duplicative records, unless the state is diligent about finding and purging them.

6. **Deceased Voters** A blunt estimate of the number of deceased people on the rolls in each state. This measure is generated through commercial records and the Social Security Administration's Death Index. If a company, such as a telemarketing firm, attempts to contact a person and is informed that the person has died, the company will make a note in its records, and the indicator may end up in Catalist's system. The measure is not comprehensive (it does not cover all deaths), nor is it as reliable as other measures (since it comes from a combination of commercial records and official records), but state-by-state differences in the number of deceased voters identified are likely to reflect the true distribution of dead voters on the rolls across jurisdictions.

7. **Deadwood** Catalist's estimate of whether a record is obsolete. A registered voter's record is considered by Catalist to be "not deadwood," "possibly deadwood, " "probably deadwood, " or "likely deadwood." The deadwood indicator takes into account the age of the registrant and if he or she is identified as deceased through commercial matches or official records. It also accounts for inactive registration status, though it is sensitive to the fact that some states have a very low threshold for placing registrants in inactive status while other states have a high threshold or do not even label any voters inactive. The measure further takes into account whether the voter has moved and has re-registered at a different address. Lastly, the deadwood model considers vote history. For instance, if a registrant has not voted even once in the last ten years, his record is unlikely to be current.

8. **Registration Date Coverage** The number for registered voters for whom their initial date of registration is unknown.

9. **Registration Distribution** The percentage of registrants who are listed with an initial registration date of January 1. January 1 is one of the least likely days of the year for a registration application to be process by an election office. Yet, many states are in the habit of assigning January 1 as a default registration date.

10. **Vote History Discrepancy** The number of discrepancies between the ballot tally and the voter tally in the 2008 and 2006 general elections. In any given election, voter turnout can be calculated in two ways: by the number of ballots cast and by the number of voters who showed up (or submitted a mail ballot). These two numbers should be exactly the same, but they rarely are. Ballots may not be counted due to machine malfunctions and voters may not be counted because an election worker might forget to check the voter's name off on the rolls.

11. **Summary Measure** The number of records that have been flagged as either deadwood or undeliverable, or both.

Before delving into a state-by-state analysis, consider some national summary statistics. Table 1 shows the initial number of all registered voters in all states plus Washington DC. It is important to note that this number is a moving target. In many states, voter files are

Table 1: National Summary Statistics

Number of Listed Registered Voters	185,445,103
Missing Address Information	244,046
Predicted Undeliverable	6,507,871
Missing Birthdate	24,501,377
Pct. Duplicative Records in Typical State	1.5%
Predicted Deceased	1,836,837
Predicted as Deadwood	6,740,264
Missing Registration Date	2,984,683
Records without Missing Address, Undeliverable, Deceased, or Deadwood Flags	169,314,778

changed multiple times a day as records are added and removed. This particular snapshot of the registration system is based on data collected from the states between March and June, 2010.

The total number, 185,445,103, reflects the sum of voters in the Catalist database at one point in time, but this number is actually already smaller than the total number of registrants listed in state and county voter files. This figure reflects registrants listed *after* Catalist has collapsed duplicative records. As we discuss in detail below, it is difficult to gauge the total number of records before duplicates are identified by Catalist, but it appears that in the typical state, about 1.5% of records are duplicates, which would mean that the true number of listed registered voters may be closer to 188 million.

Some of the quality measures we calculate cannot be identified at the level of individual voters, because they rely on distributions of grouped voters. For those that can be identified at the individual level, Table 1 lists national counts.

State-by-State Analysis

The basic purpose of voter registration lists is to record the names of citizens eligible to vote in order to authenticate voters. In most states if a citizen is not listed on the registration file prior to Election Day, he or she is not permitted to vote. Provisional ballots offer a fail safe, but the voter must still be on the lists. Registration lists serve other functions as well. Because a voter's address is listed on the file, election offices use the registration list to send important mail to voters. Voters are asked to confirm by mail that they still live at their address. They may be notified of upcoming elections or of changes to precinct locations. Mail-in ballots are sent out to addresses listed on a voter file. Furthermore, registration lists are used to identify and check voters in at precincts on an Election Day, and election workers note on the registration file whether each citizen voted or not. A registrant's recorded vote history is used, in part, as the basis for purging obsolete records from the rolls, as stipulated in the National Voter Registration Act of 1993. And of course, registration lists are used outside of administrative purposes, most notably by political campaigns seeking to mobilize

and persuade voters to vote for their candidates.

In this data presentation, we divide the administration function of voter registration lists into four broad categories. Each of the eleven quality metrics fall into one or more of these categories. The fifty states and Washington DC are evaluated on the basis of these measures.

List Function #1: Assign Voters to Precincts and Communicate with Voters by Mail

When citizens submit voter registration forms, they must be correctly assigned to their voting precincts. Because precincts are assigned to voters on the basis of their residence, proper address information is required to successfully connect a registrant to his or her voting jurisdiction. Proper address information is also required in order to contact voters by mail, which is important for notifications and mail-in ballots sent to addresses listed on the voter file. Because they both require address information, we consider these two aspects of list quality simultaneously.

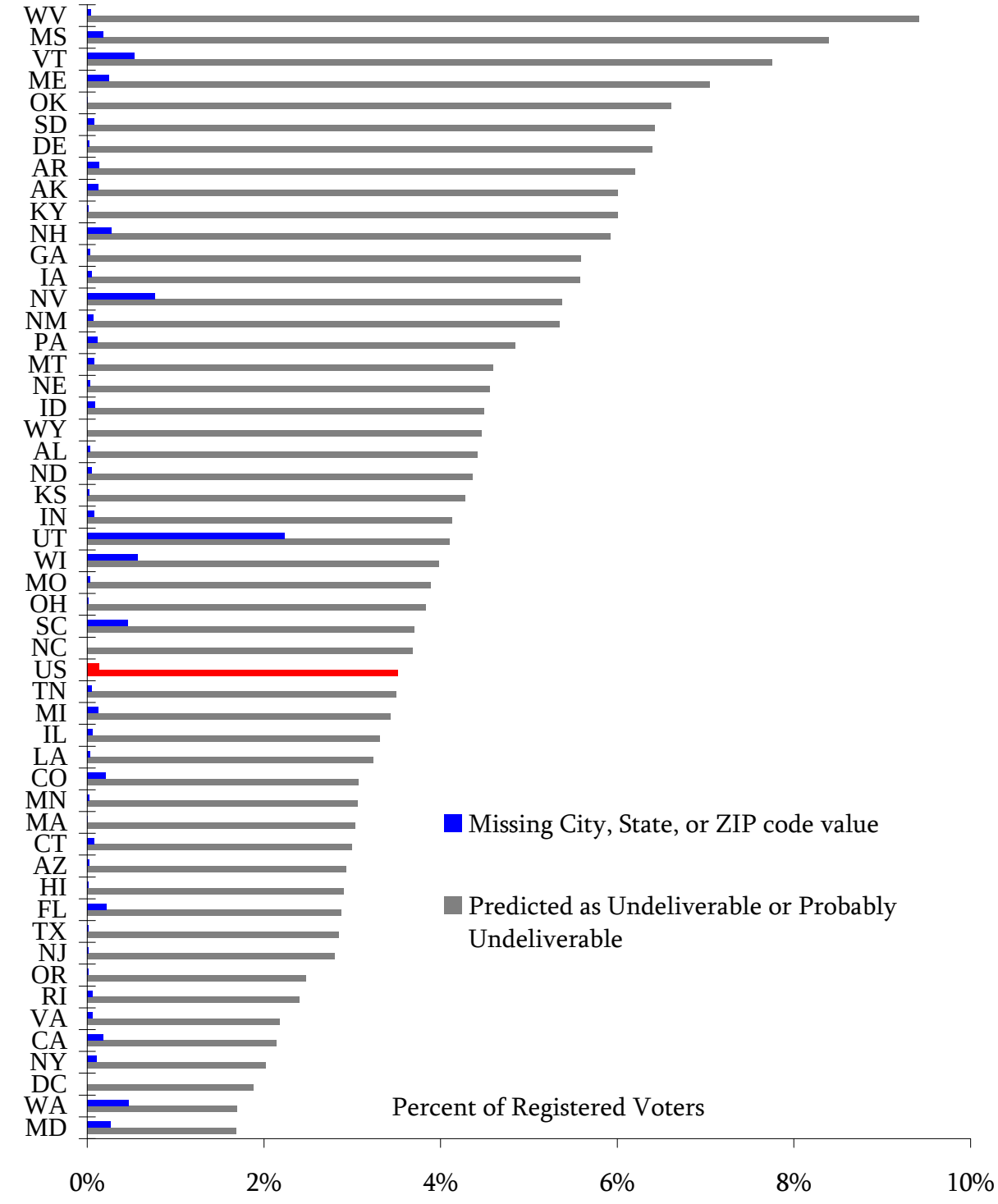
Problems with address information on registration records can take two forms: address information can be missing or address information can be incorrect. Sometimes, election offices process registration applications with incomplete address fields. They may neglect to note a voter's ZIP code, for example. Other times, due to illegible handwriting on the part of voters or faulty data entry on the part of officials, street numbers, street names, cities and ZIP codes may be entered incorrectly or contain typos.

Nationwide the incidence of incomplete addresses is 1 in 1000. That figure varies considerably across states, as depicted in Figure 1. As is evident, in most states mailing addresses have complete fields. While in general missing address fields are not pervasive, they are clearly a problem in some states. In Utah, over 1 in 50 of all registered voters have missing address information, and in neighboring Nevada, about 1 in 100 voters have missing data.

The second measure of address accuracy is a prediction of mailability, as defined above. The gray bars in Figure 1 display the rate of undeliverable or probably undeliverable records on registration lists among the states. The incidence of unmailable addresses is 1 in 25. It is clear that there is a far higher incidence of undeliverable addresses than of missing address values. In the typical state, 3.5% of registered voters have undeliverable addresses listed. Four states, Maryland, Washington State, Washington DC, and New York, have a rate as low as 2%, and fifteen states have unmailable addresses for more than 5% of the names on their lists, with West Virginia and Mississippi having the highest rates.

Three points are worth noting in considering Figure 1. First, the rate of undeliverable mail identified here does not include mail that may be undeliverable because a voter has moved residences or has died. The undeliverable mail rate depicted in Figure 1 is simply estimated by address information. Obsolete addresses are considered below, especially in Figure 9. Second, notice that these two measures of address quality are uncorrelated. This is a theme that runs throughout our exploration of these data. A state that has one problem on its lists is not more likely to have other problems. Only a handful of states do routinely well or routinely badly. Most of the statistics here, then, indicate specific problems that

Figure 1: The Accuracy of Mailing Addresses on Voter Registration Files



Source: Catalist, LLC.

must be addressed by individual states. Third, note that voters are assigned to precincts on the basis of registration addresses, whereas the two measures identified here focus on mailing addresses. While for most voters these two addresses are the same, for some they are not. The measures should be interpreted accordingly.

List Function #2: Identify Voters at the Polls

Since the 2000 election, a large number of states have strengthened their voter identification provisions, with half now requiring that voters show identification when they vote.[4] When checking in to vote at a precinct polling location on Election Day, a voter may be asked to provide proof of identification, such as a state-issued identity card with current address or a current utility bill. If the voter is to successfully match her identity to the record on the registration file, it is important that her name, address, and birthdate, as listed on the voter file, are correct. The statistics already presented in Figure 1 indicate the likely incidence of address problems. Here, we consider the likelihood of incomplete or inaccurate birthdates and flag some specific issues with the ways that state and local election offices maintain records of birthdates of voters.

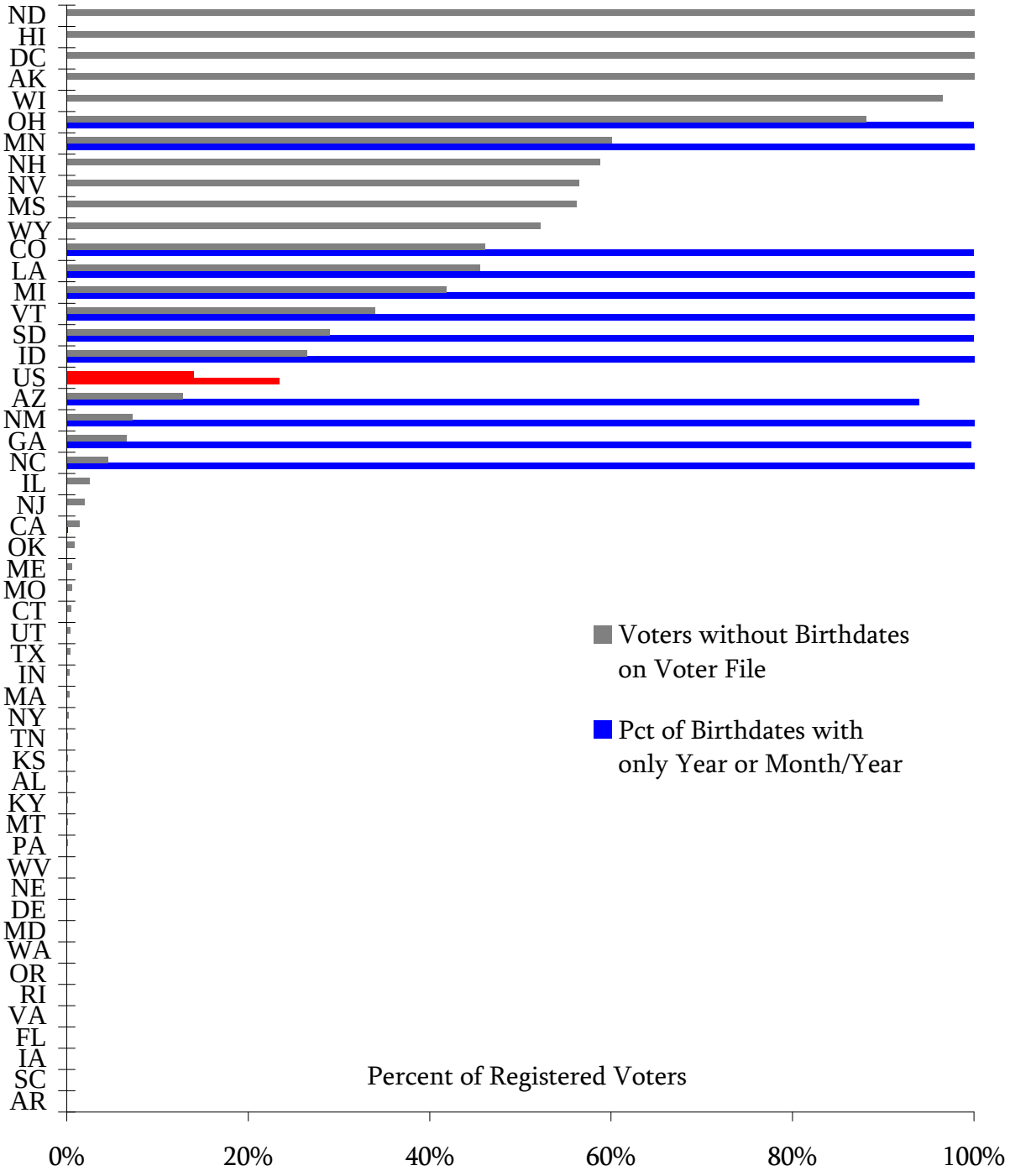
In Figure 2, we show the percentage of records for which Catalist has a record of a birthdate from the voter file. Aside from North Dakota, which does not maintain a traditional voter registration file, the states with the least amount of birthdate coverage are Alaska, Washington DC, and Hawaii. Thirteen other states have birthdates for fewer than 80% of registered voters. Most states, however, have nearly complete coverage.

As a caveat, it may be that some of the states with poorer coverage maintain birthdates for more voters than they share with data vendors, perhaps to protect the privacy of voters. However, if this were the goal, a better practice would be to follow the path of North Carolina and eleven other states in only releasing the year or month-and-year of birth for registrants. These states are noted with blue lines in Figure 2. While this practice of either collecting or sharing incomplete birthdates can protect privacy, it may interfere with the goal of confirming the identity of voters at the polls. This is a trade-off that can be taken up in future work.

Aside from birthdate coverage, we are also interested in birthdate accuracy. We can estimate the incidence of date of birth errors by looking at the distribution of birthdays on a state file. In general, if birthdates are entered correctly into the voter file, one would expect to see a uniform distribution of birthdays. That is, about the same number of people are born on January 1st as January 2nd as January 3rd, and so on. Therefore, the distribution of listed birthdays should reflect such uniformity. In Figure 3, for each month-day combination, we plot the percentage of voters whose birthdates fall on that day of the year. Voters with incomplete birthdate information and those born on February 29th are excluded. Excluding the states with either no birthdate information or with only year or month birthdate information, we are left with 35 states to analyze in Figure 3.

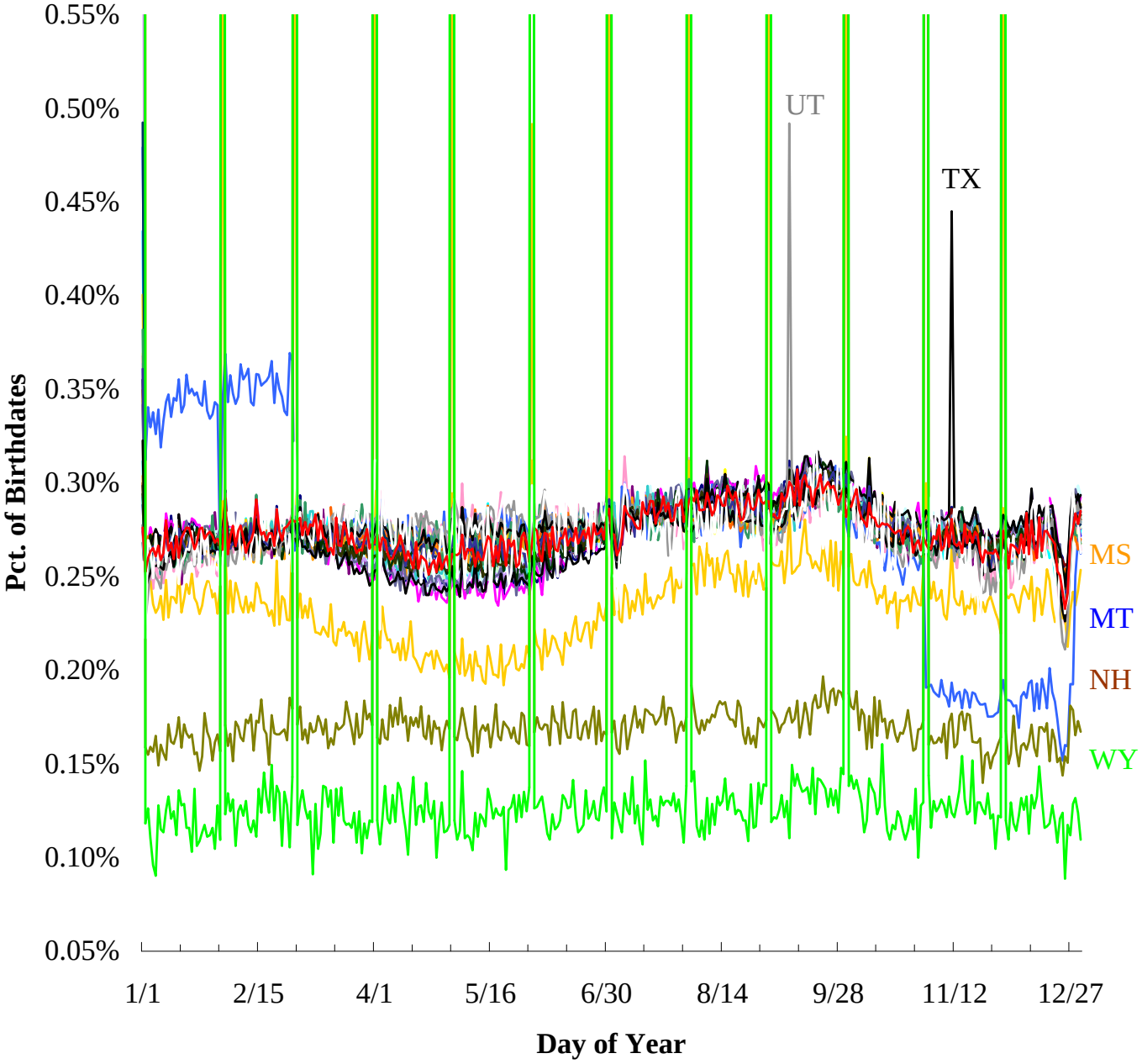
[4]For a recent summary of voter ID legislation in the states, consult the National Conference of State Legislatures website, http://www.ncsl.org/default.aspx?tabid=16602.

Figure 2: Birthdate Coverage on State Voter Files



Source: Catalist, LLC.

Figure 3: The Distribution of Voters' Birthdays on Registration Files



Source: Catalist, LLC.

As is evident from the graph, the majority of states do have voters uniformly distributed across dates. Most of the lines fluctuate around 0.27%, which is equal to 1/365. However, the exceptions stand out clearly. Mississippi, Wyoming, and New Hampshire seem to have the practice of assigning the first day of the each month to a disproportionate number of voters. Many more states have unusually high numbers of voters listed with birthdays of January 1st, a pattern that cannot be seen clearly in the Figure. Aside from Mississippi, Wyoming and New Hampshire, the states with the highest proportion of January 1st birthdays are Missouri and Pennsylvania, both of which have about twice as many January 1st birthdates as would be expected under the uniform distribution. Three other oddities that are noted in the figure are Texas, which has a uniform distribution except that a large number of voters are inexplicably listed with November 11th birthdates, Utah, which has a similar issues as Texas, and Montana, which has too many people born at the start of the year and too few at the end of the year.

List Function #3: Audits of Vote Tallies

Voter files not only contain identifying information for registered voters, but also records of past and current electoral participation. These data are used to audit elections, as they provide duplicate indicators of total numbers of voters and possible evidence of problems with machines or administration.

Vote histories have a further use in purging obsolete records. The National Voter Registration Act allows states and locales to drop from the voter registration lists any voter who has not voted in two successive federal general elections. This provision is implemented in different ways by the states, but its existence means that an election office needs accurate vote history data in order to clean the lists. If an election administrator does not keep track of which registrants voted and which did not in the last two federal elections, it cannot determine which records to purge.

Vote history accuracy can be observed by comparing the number of ballots cast in a particular jurisdiction and the number of voters listed as having voted. Official election results are tallied by counting ballots, but when voters walk into a polling place or submit an absentee ballot, their record is also marked that they voted. It is possible and quite common for the ballot tally and the registration record tally to be inconsistent in each other. A precinct election worker may neglect to indicate that a particular person voted, or a ballot may not be counted because of a machine malfunction. The first of these errors would result in a higher official tally than a voter file tally; the latter would generate the opposite result.

To measure the discrepancy between counted ballots and counted voters, we use the following procedure. For every county in a state, we take the absolute difference between the two counts. For example, if in County X, 150 ballots were counted in the 2008 general election, but only 140 registrants residing in the county were marked as having voted, we would count the deviation as 10. Similarly, if 160 voters were counted, we would also count the deviation as 10. We then sum these deviations for the entire state and divide by the total number of official votes in that election. So, a discrepancy value of 5% is interpreted as five percent of votes cast in a particular state in a particular election were inconsistent

with the tally of voters on the voter file. Figure 4 shows the discrepancy rates for the 2008 and 2006 general elections. The 2008 and 2006 vote history discrepancy rates vary considerably by state. In Oregon, North Carolina, Rhode Island, Delaware, and many other states, discrepancies are at a minimum, representing fewer than 5% of all votes. However, in other states like Mississippi, New York, and Texas, the 2008 discrepancy rate is closer to 10%. Note that the data necessary for this analysis is not available for the states of Virginia and Alaska.

Not every discrepancy represents a failure on the part of election officials, for two reasons. For one, a few states do not report the total tally of ballots as the official vote count.[5] Instead they merely report the number of ballots cast for the highest office on the ballot, usually President or Governor. Because some voters participate in down-ballot races but abstain from the top contest, this could result in a few inconsistencies between the official turnout report and the registration list count.

The second reason why some of the inconsistencies in Figure 4 do not necessarily represent an administration error is that the vote tally from a registration file excludes the votes cast by citizens who were purged from the file since the election. For instance, a person who voted in 2006 but was since removed from the rolls would not be included in the count on the registration list but would have an official ballot counted. However, this presents just a minor problem since it only applies to voters who confirmed with the registrar that they moved. Because two federal elections have not yet transpired since 2006, citizens who voted in 2006 will not have been removed without their expressed consent.

List Function #4: Maintain Current Lists

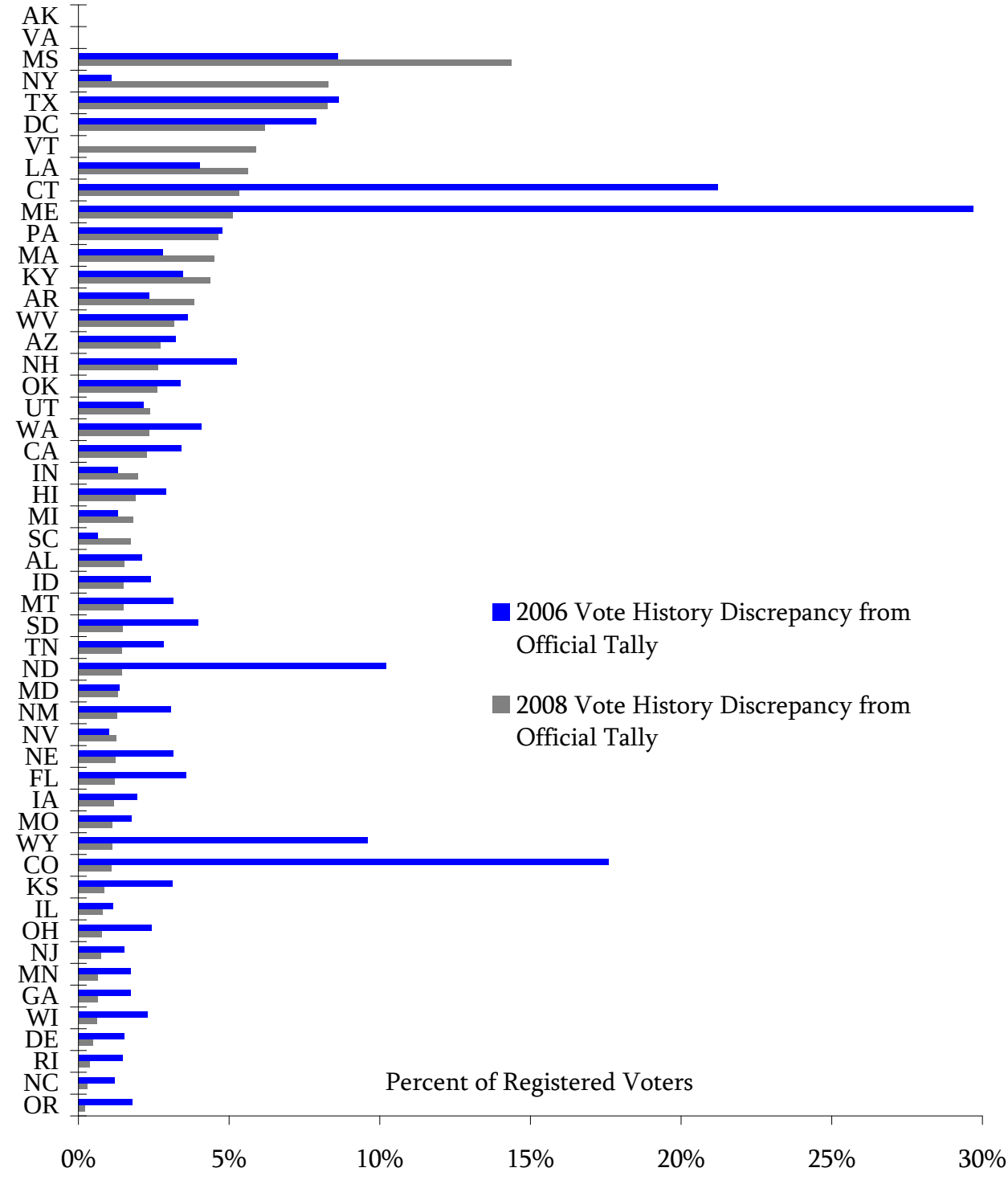
The most important function of the voter registration list is to keep a current record of those eligible to vote. This means that voters who have moved and are voting in another jurisdiction should not be on the file in their original jurisdiction. It means that deceased voters should be removed from the file in a timely fashion. It also means that voters who have not moved and are not dead and wish to be registered ought not to be purged from the voter file without their consent. Keeping current lists is the most difficult task for election administrations for the basic reason that the population is fluctuating and the lists must struggle to keep up. Because of the difficulty - and the importance - of this function of the voter list, we will focus most prominently on measuring how states are doing in this regard.

Duplicates, Deadwood, and the Dead

For our first measure of voter file currency, we asked Catalist to count the number of people on the original registration list they acquired from state and county election officials, and then to count the number they had on the list after collapsing duplicative records. When Catalist first acquires a state voter file, it searches for people who have some combination of identical names, addresses, birthdates, and state unique identifiers. If Catalist is confident

[5]For more information, consult Michael McDonald, "United States Elections Project," http://elections.gmu.edu."

Figure 4: Discrepancies between Voters Counted as Having Voted and Ballots Counted



Source: Catalist, LLC.

Table 2: Duplicates Identified on State Files

State	Total Records Acquired from Election Official	Duplicates Found	Rate of Duplicates	Official Count from State Websites	Date of Official Count	Deviation Between Official and Catalist Cnt.
MD	3,626,616	31	0.000	3,626,647	6/30/2010	0.000
TN	596,243	595	0.001	596,838	12/1/2009	0.001
NC	6,461,429	6,576	0.001	6,140,413	7/10/2010	0.052
DE	613,900	680	0.001	618,972	7/10/2010	0.008
ID	671,599	787	0.001	749,900	5/25/2010	0.104
WA	3,840,420	5,345	0.001	3,573,289	7/7/2010	0.075
NV	1,327,721	2,126	0.002	1,346,565	6/2010	0.014
WY.	259,896	506	0.002	260,818	7/1/2010	0.004
KS	1,696,065	4,275	0.003	1,698,931	6/1/2010	0.002
UT	1,480,174	4,100	0.003	1,512,433	6/17/2010	0.021
MT	637,244	2,030	0.003	668,085	11/4/2008	0.046
MN	3,129,008	10,077	0.003	3,113,664	5/1/2010	0.005
AZ	3,370,037	12,246	0.004	3,382,283	6/1/2010	0.004
RI	696,686	2,597	0.004	696,482	5/14/2010	0.000
CO	2,858,664	10,719	0.004	3,245,006	7/1/2010	0.119
AL	2,938,472	11,924	0.004	2,936,435	2010	0.001
AK	481,732	2,209	0.005	473,683	6/3/2010	0.017
NE	1,134,633	5,225	0.005	1,136,172	5/4/2010	0.001
WV.	1199297	6717	0.006	1,212,117	11/4/2008	0.011
NJ	5,234,919	35,601	0.007	5,253,112	6/4/2010	0.003
LA	2,941,151	21,207	0.007	2,913,424	7/1/2010	0.010
CA	16,841,868	122,876	0.007	16,977,031	5/24/2010	0.008
MA	4,161,453	34,352	0.008	4,220,488	10/15/2008	0.014
GA	5,746,416	48,440	0.008	5,755,750	11/4/2008	0.002
NH	922,895	8,823	0.010	922,022	5/11/2010	0.001
SC	2,871,408	30,476	0.011	2,595,131	7/12/2010	0.106
OK	2,052,290	23,069	0.011	2,038,620	1/15/2010	0.007
SD	565,137	6,445	0.011	513,203	2010	0.101
KY	2,851,937	35,908	0.013	2,861,792	6/21/2010	0.003
IA	2,252,662	36,742	0.016	2,095,890	7/1/2010	0.075
VT	447,745	7,372	0.016	444,845	7/1/2010	0.007
MS	1,956,545	32,750	0.017	1,895,583	11/4/2008	0.032
NY	11,734,910	200,016	0.017	11,658,344	4/1/2010	0.007
IL	8,411,823	166,839	0.020	7,732,908	11/4/2008	0.088
NM	1,117,474	25,530	0.023	1,131,025	5/18/2010	0.012
PA	7,477,045	200,391	0.027	8,449,679	6/28/2010	0.115
ME	812,124	41,505	0.051	853,629	6/1/2010	0.049
IN	4,538,686	313,742	0.069	4,277,762	5/4/2010	0.061
OH	7,174,313	1,218,219	0.170	8,013,556	5/4/2010	0.105
OR	2,516,751	4,150	0.002	2,036,285	7/13/2010	0.236
HI	546,298	1,220	0.002	667,647	11/8/2010	0.182
TX	9,507,837	28,204	0.003	13,023,358	3/2010	0.270
WI	4,222,057	20,156	0.005	3,469,443	7/1/2010	0.217
MO	1,956,545	32,750	0.017	4,205,774	11/4/2008	0.535
CT	224,037	3,994	0.018	2,152,320	10/27/2010	0.896
MI	8,671,755	781,057	0.090	7,470,764	11/4/2008	0.161
DC	499,082	48,094	0.096	426,531	4/30/2010	0.170
AR	2,022,685	317,926	0.157	1,622,525	6/30/2010	0.247
VA	6,995,025	1,942,752	0.278	5,005,867	6/30/2010	0.397
FL	20,389,207	7,544,552	0.370	11,085,859	5/1/2010	0.839

that the very same person is on the voter file multiple times, it collapses the duplicates into a single record. Table 2 identifies the number of records Catalist acquired from state and county elections in the Spring/Summer of 2010. It lists the number and proportion of records that are duplicative. The data in Table 2 can be misleading, however, because Catalist does

not acquire data in the same format from each state. Some states, like Florida, provide Catalist with a more raw version of their voter file than the state itself uses as its official count. As a result, the number of records Catalist acquires from Florida is very high and the number of duplicates found is also high, but this is not indicative of poor list management. On the contrary, Florida's ability to provide this raw data to Catalist suggests a superior capacity to manage data.

A different case is Connecticut, where far fewer records were acquired by Catalist than expected from the official tally of voter registrants. This may be an artifact of Catalist's procedure for updating its own lists. Rather than processing all registered voters after every acquisition from election offices, Catalist combs through new voter databases for changes (e.g. new voters, dropped voters, changed voters). Perhaps in states like Connecticut, the state provides Catalist with changes to the previous list rather than a complete new list. We are still working with Catalist to get a better understanding of the state-specific issues with calculating this measure of list quality.

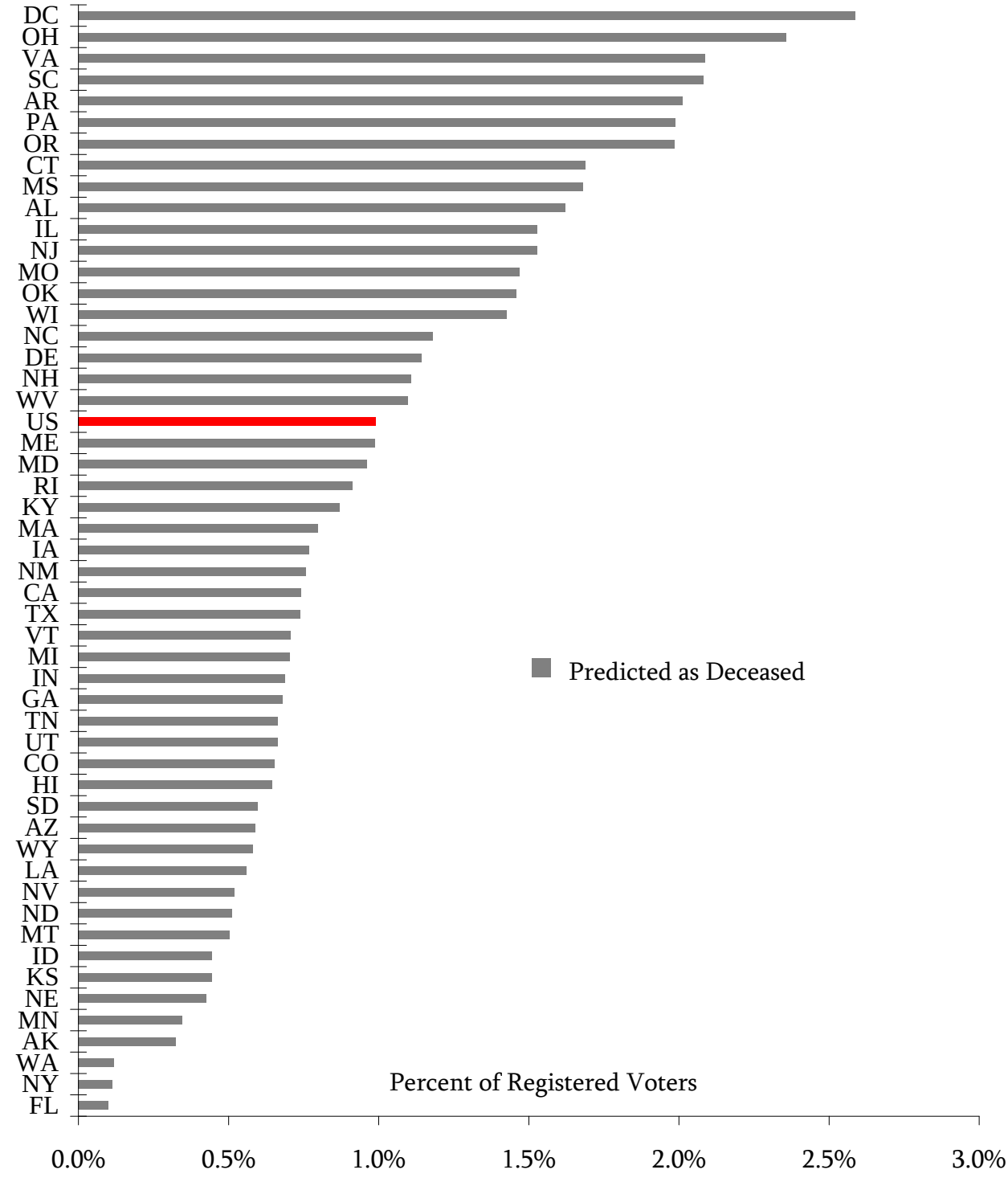
Table 2 includes official counts of registered voters that have been gathered either from state elections webpages, or in states in which registration numbers are inaccessible online from counts gathered from the 2008 general election by Michael McDonald of George Mason University.[6] States in which the deviation between Catalist's acquisition count and the official count is greater than 15% are separated and placed at the bottom of the table to emphasize that in these states particularly, the number of duplicates found by Catalist may not offer a reasonable signal of registration quality. Focusing on the states in the top of the table, the data show that while most states have very few duplicative records on their files, there is a fairly broad range in the number of duplicates found and therefore there are surely opportunities for states to improve.

While Catalist will remove duplicative records from their database during the cleansing process, they do not completely remove records of registered voters no matter how unlikely the person is to be registered. For instance, if a voter on a registration file is 120 years old or has not voted since 1972, this record is almost surely obsolete, but instead of erasing the record from its file, Catalist flags the record with a number of indicators. Figures 5 and 6 show results from two such indicators. Figure 5 displays the percentage of registered voters who from commercial or official records appear to be deceased. As with some of the other measures, while the typical state has only a fraction of a percent of deceased registrants on the rolls, a few states have quite a large number of dead registrants. In particular, in Washington DC, Ohio, Virginia, and South Carolina, over 2% of registered voters are estimated to be dead.

The overall rate of deceased voters on the file is lower than is expected given what we know about mortality rates. This might be attributable to states doing a good job at identifying and purging deceased voters, or it may be that the Catalist deceased flag does not capture a number of actual dead registrants. Even if the latter is the case, we see no reason to believe that the quality of the deceased measure should be better in some states than others. In

[6]Michael McDonald, "2008 General Election Voter Registration Statistics," *United States Elections Project*, http://elections.gmu.edu/Registration_2008G.html, accessed July 13, 2010.

Figure 5: Identified Deceased Voters on Registration Lists



Source: Catalist, LLC.

other words, this measure of deceased voters may be overall lower than in reality, but the comparison of the rate of deceased voters across states should still be valid.

Figure 6 provides yet another measure of obsolete records. Figure 6 shows the proportion of records that Catalist estimates to be definitely or probably deadwood. Values take a very wide range across states. Wyoming, which performed poorly on measures of address and birthdate accuracy performs exceedingly well in measures of deadwood. In 20 of the 51 jurisdictions, less than one percent of records look to be deadwood. An equal number of states, however, have more than 5% of their rolls consisting of deadwood, and in two states, Colorado and Arkansas, over 12% of records appear to be obsolete.

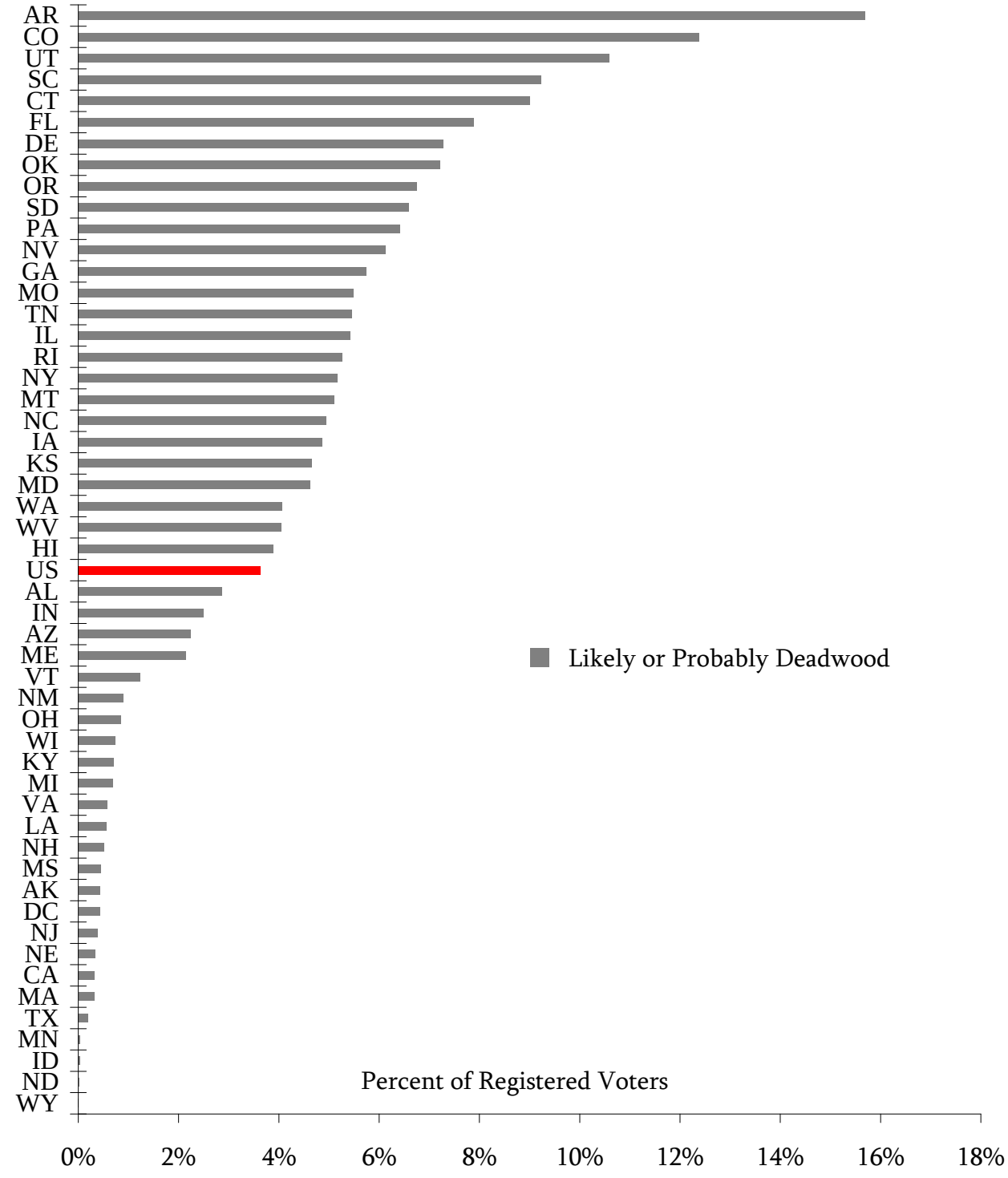
Purging Requisites

The National Voter Registration Act of 1993 sets forth guidelines to states for the purpose of removing obsolete records from the rolls. If a registrant does not confirm in writing that she is still a resident of a jurisdiction, she may only be removed from the voting file if she fails to respond to a mailed notice *and* she fails to vote in two consecutive federal elections. In order to purge old records following these guidelines, a state not only needs to have adequate mailing address information (to contact voters), adequate vote history data (as discussed above), but it also needs to keep accurate records of the length of time a voter has been on the rolls.

Accurate records of the dates on which citizens registered are important to the purging process because if an election office does not know how long a voter has been on the rolls, it cannot know whether the voter has been eligible but abstained from two consecutive federal elections. We offer two measures that assess the accuracy of registration date records. The first is the proportion of active registered voters for whom no registration date is listed. Figure 7 shows the results. The New England States of New Hampshire, Vermont, Connecticut and Maine exhibit among the highest rates of absent registration date data. In New Hampshire, nearly 40% of active registered voters do not have a registration date associated with their records. Oklahoma also has a very high rate of missing data here. Figure 7 also shows that Delaware has very few missing values, but for nearly 25% of records in that state, only a month and year are listed. Dates represented only by month and year are probably sufficient for the purging process, though the practice is clearly out of synch with the rest of the nation.

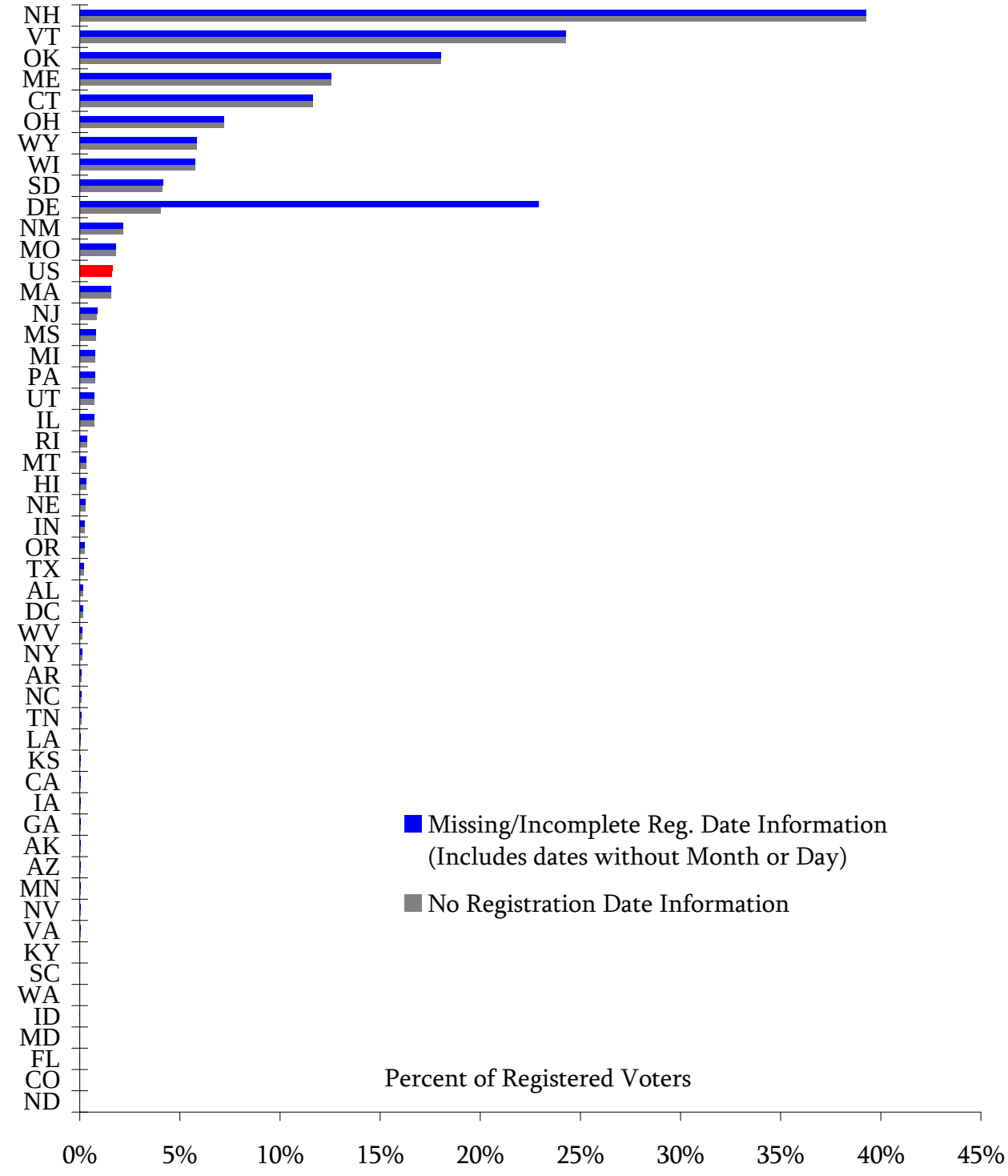
A second measure of registration date accuracy is the proportion of records in which January 1st is listed as the registration date. It appears to be the practice of registrars in many states to assign voters a date of January 1st if they did not collect or report a correct date. Of course, January 1st is one of the least likely days for a voter registration application to be processed in reality. It is a federal holiday, election offices are closed, and few jurisdictions would have registration deadlines close to the date. Nevertheless, as is shown in Figure 8, in many states an implausibly large number of voters are listed as having registered on January 1st of a particular year. In Arkansas, New York, Wisconsin and Massachusetts, over 7% of records are listed with a January 1st registration date.

Figure 6: The Proportion of Obsolete Records on Registration Files



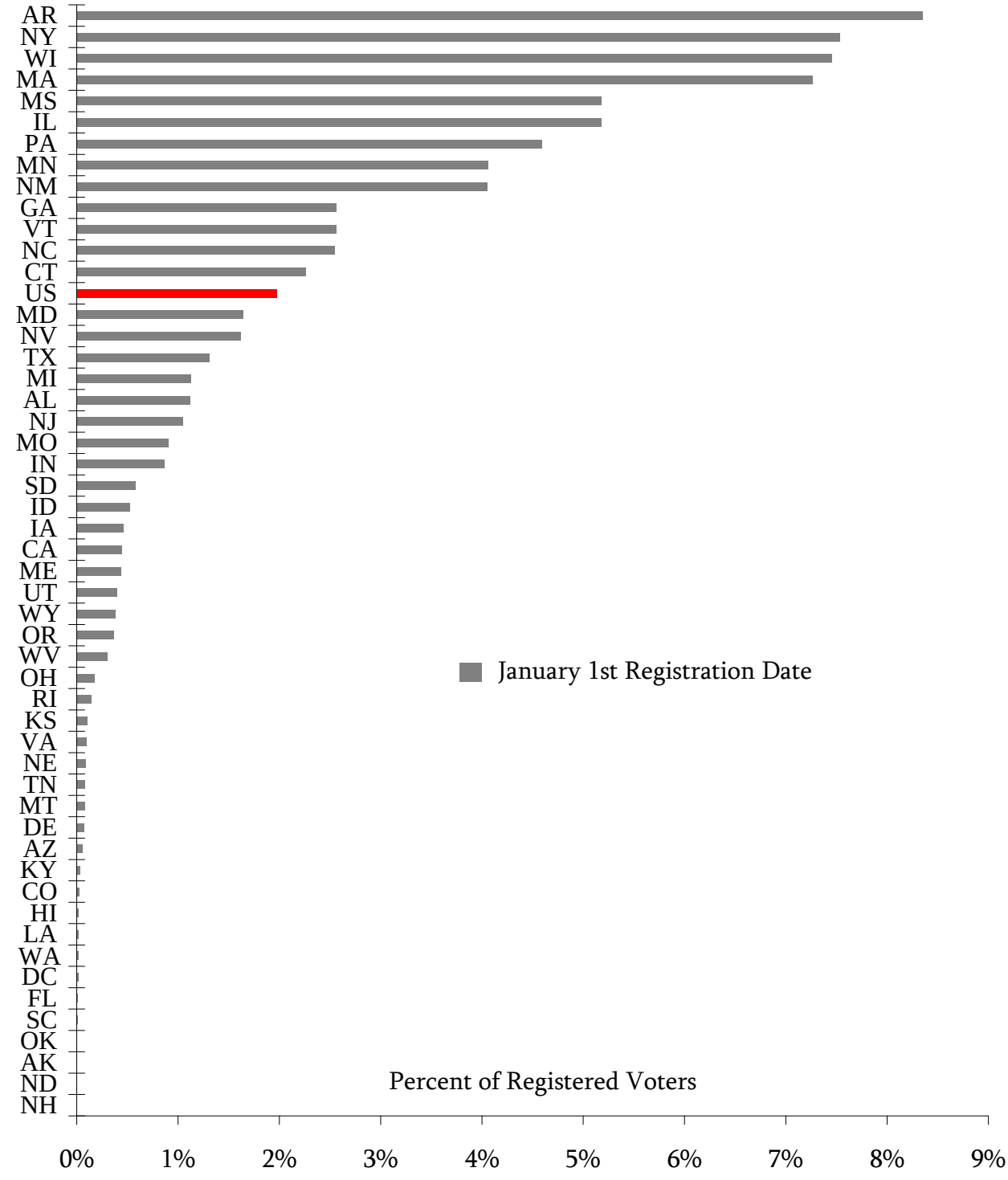
Source: Catalist, LLC.

Figure 7: Registered Voters without a Listed Registration Date



Source: Catalist, LLC.

Figure 8: Registered Voters with a Registration Date Listed as January 1st



Source: Catalist, LLC.

A Summary Index of Registration Quality

When this data project took form, we had assumed that some states would consistently perform well on these measures and other states would consistently perform poorly. The reality is different. While there are some states that generally rank well across measures, for the most part states do well on some measures and poorly on others. In statistical terms, the correlation across measures is close to zero. There seems to be no underlying "quality" parameter that is captured by averaging across these measures.

Consider a few examples. Colorado has the second highest amount of deadwood on the rolls, but it also has few dead registrants on the file and it has very good registration date information. Delaware has excellent vote history data, but has high levels of deadwood and undeliverable addresses. Texas generally ranks high, except for its vote history data and its birthdate distribution.

In Figure 9, we combine two of the more important measures described above. Figure 9 displays the percentage of registrants on a state's voter file that are either predicted as deadwood or predicted as undeliverable or both. In effect, this graph combines the data described by Figures 1 and 6. We see a very wide range of values, from 2% of records appearing problematic in Washington, DC to 20% in Arkansas.

Using a very different methodological approach, in a recent study supported by the Pew Center on the States we found results consistent with those on display in Figure 9. Along with our colleagues Alan Gerber and David Doherty at Yale University, we audited the registration files in the state of Florida and the county of Los Angeles California and measured the rate of mail returned as undeliverable.[7] Mail is undeliverable either because the address is incorrect (missing data or typos) or because the registrant is no longer alive or no longer at the address. Thus, the rate of undeliverable mail in that study should track with the data in Figure 9. In Florida we found an undeliverable rate of 9% (compared to 10% here) and in Los Angeles we found a rate of 5%, compared to 2.5% for California in this study.

Figure 9 summarizes two important scores, but because these two scores are uncorrelated with each other, the combined score obscures more than it reveals. Colorado is a case in point. In the combined measure in Figure 9, Colorado appears as the second worst state. This ranking obscures the fact that Colorado does better then three fifths of the states on maintaining mailable addresses. Colorado's low score relative to the other states is attributable to its poor ranking on only one of these measures, not on both.

Moving Forward

The purpose of this report is to identify a new way of measuring the quality of elections in the United States. By collaborating with a commercial data vendor, we have gained important insights into voter registration data that are at the core of election administration.

[7] Stephen Ansolabehere, David Doherty, Alan Gerber, and Eitan Hersh, "Voter Registration List Quality Pilot Studies," *Pew Center on the States*, 2010.

Figure 9: Registration Records Estimated as Undeliverable or Deadwood, or Both

Source: Catalist, LLC.

The measures we have described and estimated are not the only ways to assess quality, but they provide a broad and comprehensive perspective into the degree and variation of problems in voter registration files across the country.

On each of nearly a dozen items, we have ranked the states in terms of the quality of their records. Every measure reveals a range of scores, and it will surely be useful for states to know where they stand relative to one another. The measures can help an election office identify areas most in need of improvement. For national policymakers and advocates, these initial results suggest that the path forward should focus on addressing each state's unique set of weak spots rather than identifying states that are overall delinquent or by proposing broad changes to election administration. It does not seem to be the case that there are good states and bad states, and that the bad states can be shamed into emulating the good states.

What is more, the states that leave control over election administration primarily in the hands of county officials do not seem to perform worse on these measures than states with centralized systems. California is known to have the most "bottom-up" registration system in the nation. There, the state election office does not input changes or updates to the files originating from the counties. And yet, California performs well above average on these measures of registration quality. At the same time, in preliminary county-by-county analyses we have conducted, we find that, in general, rural counties have a greater incidence of obsolete records than urban counties.

Apart from comparing the quality of records across jurisdictions, these quality measures also offer a sense of the national average of registration data quality. Interpretations of how well or how poorly the nation is doing as a whole will differ depending on one's perspective. Our own view as informed by the data presented here is that overall the states are doing quite well. Indeed, the greatest barrier to maintaining clean election records are beyond the control of state and local governments; that voters move, die, enter and exit the registration system means that keeping a list of eligible voters is a challenge. Given the population fluctuations, we see the rate of bad addresses, deadwood, duplicates, and dead voters to be reasonable and actually lower than expected. In the areas in which election administrators can work to make improvements, such as maintaining complete addresses, birthdates, registration dates, and vote histories, there are certainly opportunities for states to do better and for voters and advocates to demand better. We hope that these measures shed light on opportunities for reform.

We welcome the feedback of policymakers and colleagues to improve upon the work we have begun here.