A Appendix

A.1 Modeling the Birthday Distribution

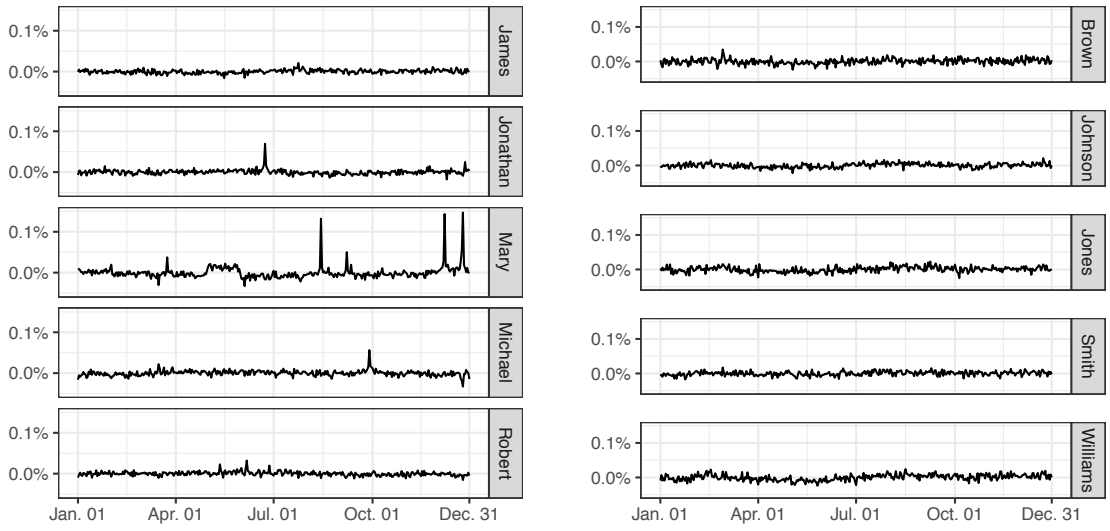
Our goal is to estimate $\Pr(B = b \mid F = f, L = l, Y = y)$, the probability that a voter has a birthday b conditional on having first name f, last name l, and being born in year y. The challenge is that we do not observe a sufficient number of people with the same name who were born in the same year to estimate this only using the empirical distribution. Our first simplification is to assume that $\Pr(B = b \mid F = f, L = l, Y = y) = \Pr(B = b \mid F = f, Y = y)$, so that we can ignore an individual's last name when estimating this probability. The justification for this assumption comes from Figure A.1, which plots the difference in the share of voters with the most common first and last names born on a given day and the share of the general population of voters born on that same day. The left panel of the plot shows a disproportionate number of voters named John and Mary are born on St. John's Day (June 24) and near Christmas, respectively. The right panel does not show similar spikes in the common last names. This pattern is understandable since first names are actively selected whereas last names are generally not. Proposition 1 derives our estimate of $\Pr(B = b \mid F = f, Y = y)$ under three assumptions.

Proposition 1. *Assume:*

1. *If* $d_{b,y_1} = d_{b,y_2} \forall b$, *then* $\Pr(B = b \mid Y = y_1, F = f) = \Pr(B = b \mid Y = y_2, F = f)$;
2. $\Pr(F = f, D = d \mid B = b) = \Pr(F = f \mid B = b) \Pr(D = d \mid B = b)$;
3. $\Pr(D = d \mid B = b) = \Pr(D = d)$.

Then we have,

$$\Pr(B = b \mid F = f, Y = y) = \frac{Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{\sum_{b'} \Pr(B = b' \mid F = f)\ \Pr(D = d_{b',y})}\ . \tag{8}$$



(a) (b)

Figure A.1: Examples of names among 2012 voters with a non-uniform date of birth distribution, by day (a) or year (b) of birth.

The first assumption means that if y_1 and y_2 are two different years with the same weekday schedule, then the distribution of birthdays for a given first name is the same. Two years have the same weekday schedule when January 1st falls on the same day of the week in both years, and neither or both years are a leap year. Note that while this assumption means that someone named Connor born in 1973 would have the same probability of being born on January 1st as someone named Connor born in 1979, as both were Mondays, it does not require the number of Connors born in 1973 and 1979 to be the same. We use the notation $y' \sim y$ to indicate that year y' has the same weekday schedule as year y.

The second assumption means that the distribution of first names of people born on a given day is independent of the day of the week. So once we condition on being born on a given day, nothing is learned about what day of the week one was born on from one's first name. While we acknowledge there are cases — like being named Wednesday or Domingo — where this assumption is not correct, such cases are relatively rare.

The third assumption is that birthday and birth day-of-week are independent. Thus, knowing an individual's birthday does not give us any information on the day of the week

they were born on.

Proof of Proposition 1

Consider the set of people born with first name f and birthday b on day of the week $d_{b,y}$, which is represented by $\{B = b, D = d_{b,y}, F = f\}$. Without loss of generality, we can decompose this set into the union of sets of people born with first name f and birthday b in a year y' such that $d_{b,y'} = d_{b,y}$. Going one step further, and ignoring leap years, we can say that $d_{b,y'} = d_{b,y}$ is equivalent to y' and y having the same weekday schedule, which we can write as $y' \sim y$ using our notation:

$$\{B = b, D = d_{b,y}, F = f\} = \bigcup_{(y' \text{ s.t. } y' \sim y)} \{B = b, Y = y', F = f\}.$$

Because the sets on the right-hand side of the equation above correspond to different years, and thus have no intersection, we can write,

$$\begin{aligned}
\Pr(B = b, D = d_{b,y}, F = f) &= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b, Y = y', F = f), \\
\Pr(B = b, D = d_{b,y} \mid F = f)\ \Pr(F = f) &= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b, Y = y' \mid F = f)\ \Pr(F = f), \\
\Pr(B = b, D = d_{b,y} \mid F = f) &= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b, Y = y' \mid F = f) \\
&= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b \mid Y = y', F = f)\ \Pr(Y = y' \mid F = f).
\end{aligned}$$

Assumption 1 gives us that $\forall y' \sim y, \Pr(B = b \mid Y = y', F = f) = \Pr(B = b \mid Y = y, F = f))$, so that,

$$\begin{aligned}
\Pr(B = b, D = d_{b,y} \mid F = f) &= \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(B = b \mid Y = y', F = f)\ \Pr(Y = y' \mid F = f) \\
&= \Pr(B = b \mid Y = y, F = f) \sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f).
\end{aligned}$$

Rearranging terms, we get,

$$\Pr(B = b \mid Y = y, F = f) = \frac{\Pr(B = b, D = d_{b,y} \mid F = f)}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)}. \tag{9}$$

Using Bayes' rule, we can rewrite the numerator in Eq. (9) as,

$$\begin{aligned} \Pr(B = b, D = d_{b,y} \mid F = f) &= \frac{\Pr(F = f, D = d_{b,y} \mid B = b)\ \Pr(B = b)}{\Pr(F = f)} \\ &= \frac{\Pr(F = f \mid B = b)\ \Pr(D = d_{b,y} \mid B = b)\ \Pr(B = b)}{\Pr(F = f)} \end{aligned} \tag{10}$$

where the second equality comes from assumption 2, which gives us that $\Pr(F = f, D = d \mid B = b) = \Pr(F = f \mid B = b) \Pr(D = d \mid B = b)$. By Bayes' rule,

$$\Pr(F = f \mid B = b) = \frac{\Pr(B = b \mid F = f)\ \Pr(F = f)}{\Pr(B = b)}. \tag{11}$$

Plugging Eq. (11) into Eq. (10) and simplifying gives us that

$$\begin{aligned} \Pr(B = b, D = d_{b,y} \mid F = f) &= \Pr(F = f \mid B = b) \times \Pr(D = d_{b,y} \mid B = b) \times \frac{\Pr(B = b)}{\Pr(F = f)} \\ &= \frac{\Pr(B = b \mid F = f)\ \Pr(F = f)}{\Pr(B = b)} \times \Pr(D = d_{b,y} \mid B = b) \times \frac{\Pr(B = b)}{\Pr(F = f)} \\ &= \Pr(B = b \mid F = f)\ \Pr(D = d_{b,y} \mid B = b) \\ &= \Pr(B = b \mid F = f)\ \Pr(D = d_{b,y}) \end{aligned} \tag{12}$$

where the final equality comes from assumption 3, which gives us that $\Pr(D = d \mid B = b) = \Pr(D = d)$. Substituting the results of Eq. (12) into the numerator of Eq. (9) gives us that

$$\begin{aligned}\Pr(B = b \mid Y = y, F = f) &= \frac{\Pr(B = b, D = d_{b,y} \mid F = f)}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)} \\ &= \frac{\Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{\sum_{(y' \text{ s.t. } y' \sim y)} \Pr(Y = y' \mid F = f)} \\ &= \frac{\Pr(B = b \mid F = f)\ \Pr(D = d_{b,y})}{Z(f, y)}. \end{aligned} \tag{13}$$

To solve for $Z(f, y)$ we note that it must be the case that $\sum_{b'} \Pr(B = b' \mid Y = y, F = f) = 1$ for it to be a valid probability distribution. Thus,

$$Z(f, y) = \sum_{b'} \Pr(B = b' \mid F = f)\ \Pr(D = d_{b',y}). \tag{14}$$

Plugging in Eq. (14) to Eq. (13) yields the proposition.

□

A.2 Statement and Proof of Theorem 1

Theorem 1. *Suppose $D_{f,l,y}$ is a discrete probability distribution of birthdays $b_1, \ldots, b_n$ with $\Pr_{D_{f,l,y}}(b_i) = p_{b_i|f,l,y}$. Further assume there are $q \geq 1$ independent observations from $D_{f,l,y}$, $B_1, \ldots, B_q$, and $k_{f,l,y} \leq q$ copies $B_{q+1}, \ldots, B_{q+k_{f,l,y}}$ such that $B_{q+i} = B_i$. Let $M_{f,l,y}$ be the number of pairwise matches among the $n_{f,l,y} = q + k_{f,l,y}$ observations, and define the estimator*

$$\hat{k}_{f,l,y} = \left(M_{f,l,y} - \binom{n_{f,l,y}}{2} \sum_i p^2_{b_i|f,l,y} \right) \Bigg/ \left(1 - \sum_i p^2_{b_i|f,l,y} \right). \tag{15}$$

Then $\mathbb{E}\hat{k}_{f,l,y} = k_{f,l,y}$ and

$$\operatorname{Var}(\hat{k}_{f,l,y}) \leq 4\binom{n_{f,l,y}}{2} \left[\frac{\sum_i p^2_{b_i|f,l,y}}{1 - \sum_i p^2_{b_i|f,l,y}} \right] + 12\binom{n_{f,l,y}}{3} \left[\frac{\sum_i p^3_{b_i|f,l,y} - \left(\sum_i p^2_{b_i|f,l,y}\right)^2}{\left(1 - \sum_i p^2_{b_i|f,l,y}\right)^2} \right].$$

Proof. To simplify the notation, we represent $M_{f,l,y}$ by M, $n_{f,l,y}$ by n, $D_{f,l,y}$ by D, $p_{b_s|f,l,y}$ by p_s, and $k_{f,l,y}$ by k. We start by computing the expectation of M. For $1 \leq i < j \leq q+k$, let $A_{i,j}$ indicate whether $B_i = B_j$. Then by the linearity of expectation,

$$\mathbb{E}M = \mathbb{E}\left(\sum_{1\leq i<j\leq q+k} A_{i,j}\right) = \sum_{1\leq i<j\leq q+k} \mathbb{E}A_{i,j}. \tag{16}$$

For $1 \leq i \leq k$, $\mathbb{E}A_{i,q+i} = 1$ since $B_i = B_{q+i}$ by construction. For the remaining $\binom{q+k}{2} - k$ terms, $\mathbb{E}A_{i,j} = \Pr_D(B_i = B_j) = \sum_s p_s^2$. Consequently,

$$\begin{aligned}
\mathbb{E}M &= k + \left(\binom{q+k}{2} - k\right)\sum_s p_s^2 \\
&= k\left(1 - \sum_s p_s^2\right) + \binom{q+k}{2}\sum_s p_s^2.
\end{aligned}$$

By rearranging terms, we now have that $\mathbb{E}\hat{k} = k$.

To compute the variance of $\hat{k}$, we first compute the variance of M, decomposing it as

$$\mathrm{Var}(M) = \sum_{1\leq i<j\leq q+k} \mathrm{Var}(A_{i,j}) + 2\sum_{\mathcal{R}} \mathrm{Cov}(A_{i,j}, A_{k,l}) \tag{17}$$

where $\mathcal{R}$ is the set of indices so that each distinct, unordered pair $(A_{i,j}, A_{k,l})$ appears in the sum exactly once. Since $A_{i,j}$ is an indicator variable,

$$\mathrm{Var}(A_{i,j}) = \mathbb{E}A_{i,j} - \left(\mathbb{E}A_{i,j}\right)^2. \tag{18}$$

By the above, $\mathrm{Var}(A_{i,q+i}) = 0$ for $1 \leq i \leq k$; and for the remaining terms, $\mathrm{Var}(A_{i,j}) = \sum_s p_s^2 - \left(\sum_s p_s^2\right)^2$. Consequently,

$$\sum_{1\leq i<j\leq q+k} \mathrm{Var}(A_{i,j}) = \left(\binom{q+k}{2} - k\right)\left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{19}$$

Next we consider the covariance terms $\mathrm{Cov}(A_{i,j}, A_{k,l})$, dividing them into two sets and analyzing them separately.

Case 1: We first consider the terms where the indices i, j, k, l are all distinct. If neither B_i nor B_j are copies of either B_k or B_l, then $A_{i,j}$ and $A_{k,l}$ are clearly independent, and so $\mathrm{Cov}(A_{i,j}, A_{k,l}) = 0$. Now suppose that exactly one (but not both) of $\{B_i, B_j\}$ is a copy of either B_k or B_l. In this case, since each observation can be a copy of at most one other observation, B_i cannot be a copy of B_j, and B_k cannot be a copy of B_l. We thus have,

$$\mathbb{E}A_{i,j} = \mathbb{E}A_{k,l} = \sum_s p_s^2 \qquad \text{and} \qquad \mathbb{E}A_{i,j}A_{k,l} = \sum_s p_s^3.$$

Consequently,

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) = \sum_s p_s^3 - \left(\sum_s p_s^2\right)^2.$$

Moreover, there are $2k\left[\binom{q+k-2}{2} - (k-1)\right]$ such instances where there is a single copy between $\{B_i, B_j\}$ and $\{B_k, B_l\}$. To see this, note that we can enumerate the instances by first selecting one of the k copies (and its pair); then selecting two additional observations from the remaining $q+k-2$ while avoiding the $k-1$ combinations that result in selecting another copy and its pair; and lastly, choosing one of the two ways in which the selected observations can be combined to form two unordered pairs.

Finally, suppose that both B_i and B_j are copies of B_k and B_l. As above, B_i cannot be a copy of B_j, and B_k cannot be a copy of B_l, so

$$\mathbb{E}A_{i,j} = \mathbb{E}A_{k,l} = \sum_s p_s^2 \qquad \text{and} \qquad \mathbb{E}A_{i,j}A_{k,l} = \sum_s p_s^2.$$

Consequently,

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) = \sum_s p_s^2 - \left(\sum_s p_s^2\right)^2.$$

There are $2\binom{k}{2}$ such terms, since we must first select two of the k copies, and then select one

of the two ways in which to combine the four random variables into two unordered pairs.

Case 2: We next consider the covariance terms where there are three distinct indices among the set $\{i, j, k, l\}$. Since $i \neq j$ and $k \neq l$, this means that $\{i, j\} \cap \{k, l\} \neq \emptyset$. If there are no copies among the three distinct random variables, then

$$\mathbb{E}A_{i,j} = \mathbb{E}A_{k,l} = \sum_s p_s^2 \qquad \text{and} \qquad \mathbb{E}A_{i,j}A_{k,l} = \sum_s p_s^3$$

and so,

$$\text{Cov}(A_{i,j}, A_{k,l}) = \sum_s p_s^3 - \left(\sum_s p_s^2\right)^2.$$

The number of such terms—with three distinct random variables, none of which are copies of one another—is $3\left[\binom{q+k}{3} - k(q+k-2)\right]$. To count the terms, we first count the $\binom{q+k}{3}$ ways of selecting three variables from the $q+k$, and then subtract the number of possibilities in which one variable is a copy of another. This latter quantity can be obtained by first selecting one of the k copied variables and its pair, and then selecting a third observation from the remaining $q+k-2$. Finally, given the three random variables, we form two pairs by selecting which one of the three to duplicate, and replicating that selected variable in each pair.

Now, if B_i is a copy of B_j, then $A_{i,j} = 1$. Consequently, $A_{i,j}$ and $A_{k,l}$ are independent, and so $\text{Cov}(A_{i,j}, A_{k,l}) = 0$. An analogous argument holds if B_k is a copy of B_l.

Finally, if the non-repeated variable among $\{B_i, B_j\}$ is a copy of the non-repeated variable among $\{B_k, B_l\}$, then

$$\mathbb{E}A_{i,j} = \mathbb{E}A_{k,l} = \sum_s p_s^2 \qquad \text{and} \qquad \mathbb{E}A_{i,j}A_{k,l} = \sum_s p_s^2$$

and so,

$$\text{Cov}(A_{i,j}, A_{k,l}) = \sum_s p_s^2 - \left(\sum_s p_s^2\right)^2.$$

Such terms number $k(q+k-2)$, since we must select a copied random variable and its pair,

and then a third random variable among the remaining $q+k-2$ to replicate.

Aggregating all the above terms, we have,

$$\begin{aligned}\mathrm{Var}(M) &= \left[\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right]\left[\binom{q+k}{2} - k + 4\binom{k}{2} + 2k(q+k-2)\right] \\ &\quad + \left[\sum_s p_s^3 - \left(\sum_s p_s^2\right)^2\right]\left[4k\binom{q+k-2}{2} - 4k(k-1) + 6\binom{q+k}{3} - 6k(q+k-2)\right].\end{aligned}$$

Since $\mathrm{Var}(\hat{k}) = \mathrm{Var}(M)/\left(1-\sum_s p_s^2\right)^2$,

$$\begin{aligned}\mathrm{Var}(\hat{k}) &= \left[\frac{\sum_s p_s^2}{1-\sum_s p_s^2}\right]\left[\binom{q+k}{2} + 4\binom{k}{2} + 2k(q+k-2) - k\right] \\ &\quad + \left[\frac{\sum_s p_s^3 - \left(\sum_s p_s^2\right)^2}{\left(1-\sum_s p_s^2\right)^2}\right]\left[4k\binom{q+k-2}{2} + 6\binom{q+k}{3} - 4k(k-1) - 6k(q+k-2)\right].\end{aligned}$$

Finally, to derive an upper bound on $\mathrm{Var}(\hat{k})$ that is independent of k, observe that $\sum_s p_s^2 \leq \sum_s p_s = 1$, and so $\sum_s p_s^2/(1-\sum_s p_s^2) \geq 0$. Moreover, by Jensen's inequality applied to the convex function $\phi(x) = x^2$ and weights p_i, $\sum_s p_s^3 \geq (\sum_s p_s^2)^2$. Thus, the two terms involving p_i in the variance expression above are non-negative. Consequently, dropping the negative terms, and noting that $k \leq (q+k)/2$, we get the bound

$$\mathrm{Var}(\hat{k}) \leq 4\binom{q+k}{2}\left[\frac{\sum_s p_s^2}{1-\sum_s p_s^2}\right] + 12\binom{q+k}{3}\left[\frac{\sum_s p_s^3 - \left(\sum_s p_s^2\right)^2}{\left(1-\sum_s p_s^2\right)^2}\right].$$

On the other hand, to derive a lower bound, we can minimize positive terms and maximize negative terms in the variance expression. Considering $k \leq (q+k)/2$, observe that $4\binom{k}{2} + 2k(q+k-2) - k \geq -\frac{q+k}{2}$, and $4k\binom{q+k-2}{2} - 4k(k-1) - 6k(q+k-2) \geq -4(\frac{q+k}{2})(\frac{q+k}{2}-1) -$

$6(\frac{q+k}{2})(q+k-2) = -4(q+k)(q+k-2)$. So we can write

$$\begin{aligned}\mathrm{Var}(\hat{k}) \;\geq\; & \left[\binom{q+k}{2} - \frac{q+k}{2}\right]\left[\frac{\sum_s p_s^2}{1-\sum_s p_s^2}\right] \\ & + \left[6\binom{q+k}{3} - 4(q+k)(q+k-2)\right]\left[\frac{\sum_s p_s^3 - \left(\sum_s p_s^2\right)^2}{\left(1-\sum_s p_s^2\right)^2}\right].\end{aligned}$$

□

A.3 Statement and Proof of Proposition 2

Proposition 2. *Assume a set of $n \geq 1$ objects, out of which k^{orig} objects are duplicates, and the rest are unique. Additionally assume that each object has at most one duplicate in the set. Then suppose that each one of these n objects is copied with probability p_u, and dropped from the set with probability p_r. Assume K to be the number of unique objects with a copy in the updated set, and N to be the size of this set. If we define the estimator $\hat{k}^{orig}$ as,*

$$\hat{k}^{orig} = \frac{K}{(1-p_r)^2 - 2p_u} - \frac{Np_u}{(1+p_u-p_r+p_up_r)\left((1-p_r)^2-2p_u\right)} \tag{20}$$

then $\mathbb{E}\hat{k}^{orig} = k^{orig}$.

Proof. We start by computing the expectation of K. By definition, K is the number of unique objects with a copy observed in the updated set. Initially and before updating the set, there are $n-k^{\mathrm{orig}}$ unique objects out of which k^{orig} objects have a copy in the set, and the remaining $n-2k^{\mathrm{orig}}$ objects are with no duplicates. Each of these k^{orig} objects will still have a copy in the updated set if and only if neither itself nor its copy is dropped. The probability that an object and its copy are not dropped is $(1-p_r)^2$. For the remaining $n-2k^{\mathrm{orig}}$ unique objects, each will have copy in the updated set if and only if it gets duplicated, which has a probability of p_u. Therefore,

$$\mathbb{E}K = k^{\mathrm{orig}}(1-p_r)^2 + (n-2k^{\mathrm{orig}})p_u = k^{\mathrm{orig}}\left[(1-p_r)^2 - 2p_u\right] + np_u. \tag{21}$$

Rearranging terms, we get,

$$\mathbb{E}\left[\frac{K - np_u}{(1-p_r)^2 - 2p_u}\right] = k^{\text{orig}}. \tag{22}$$

n is the number of objects in the original set, while N is the size of updated set. Each object in the original set contributes two objects to the updated set with probability p_u, or one object with probability $(1-p_u)(1-p_r) = 1 - p_u - p_r + p_u p_r$. Therefore,

$$\mathbb{E}N = \sum_{i=1}^{n} 2p_u + 1 - p_u - p_r + p_u p_r = n(1 + p_u - p_r + p_u p_r) \tag{23}$$

Substituting $n = \frac{\mathbb{E}N}{1+p_u-p_r+p_u p_r}$ into the Eq. (22), we have $\mathbb{E}\hat{k}^{\text{orig}} = k^{\text{orig}}$.

Note that in the proof of Theorem 1 we were estimating the number of pairs of duplicates in the set, while here we are interested in the number of unique records with duplicates in the set. As long as we assume a person does not vote more than twice in the election, the two estimation approaches yield the same result. □

A.4 Name and DOB Errors in the Voter File

To estimate the number of people who voted twice in the 2012 election, we use Target Smart's national voter file, which lists the first name, middle name,[18] last name, suffix, date of birth, and turnout history associated with a voter registration.[19] These data provide a nearly comprehensive list of 2012 general election participation: the data include 126,414,090 vote records from the 2012 election, as compared to the 129,085,410 votes cast for a presidential

[18] Although the data include middle name, we do not use this information in our analysis. First, states do not require middle name to be reported and not everyone has a middle name. Among those who both have a middle name and report it, the information is often recorded inconsistently. Many records also contain only a middle initial, making it difficult to assess the accuracy of a given match. Other records have what appear to be transcription errors, such as a suffix in the middle name field.

[19] Some states do not reveal the full date of birth on each registration. In such cases, Target Smart supplements the missing birthdates with information obtained from commercial data sources.

candidate nationwide.[20] 124,942,823 of these 126,414,090 vote records have a non-missing first name, last name, and DOB. Before using the data, we standardize first names in the voter file by converting nicknames to their canonical form. We use pdNickname software, which contains tables relating nicknames to canonical names. We only consider short form or diminutive nicknames with the highest relationship quality scores (less than 5). If a nickname maps to multiple canonical names, we convert it to the most popular canonical name among voters with the same gender. For instance, a male voter named Chris is considered Christopher, and a female voter named Chris is considered Christine.

One concern with these data is that date of birth may not always be reported accurately in the voter file. Figure A.2 shows the distribution of birthdays (i.e., month and day of birth) for voter registrations with a birth year of 1970 and a vote record in 2012. It illustrates a pattern, also shown by Ansolabehere and Hersh (2010), that too many registration records indicate that a voter was born on first day of the month. Across all years, about 14% of 2012 vote records are indicated to have been born on the first day of the month.[21] Such measurement error could cause us to incorrectly count two votes cast by distinct voters as instead coming from a single voter, and thus overestimate the true rate of double voting.

We also suspect that the birthdates of individuals in multi-generational households are reported incorrectly in a few states. When we match vote records within states by not only first name, last name, and date of birth, but also registration address, we find 7,504 and 2,350 in-state duplicate voters in Mississippi and Wisconsin, respectively. In a vast majority of these cases, the records share a different middle name or suffix, suggesting a situation in which either a father (mother) or son (daughter) were assigned the others' birthdate. Figure A.3 shows the distribution of potential multi-generational matches within states, normalized based on the size of the state. In addition to Wisconsin and Mississippi, we

[20] http://www.fec.gov/pubrec/fe2012/federalelections2012.pdf

[21] We can detect some other improbable clumps of birthdays in a few states. For instance, March 26th in Wisconsin and New Hampshire, June 5th in Idaho, and the whole month of January in Hawaii all show a higher concentration of certain voter registration birthdays

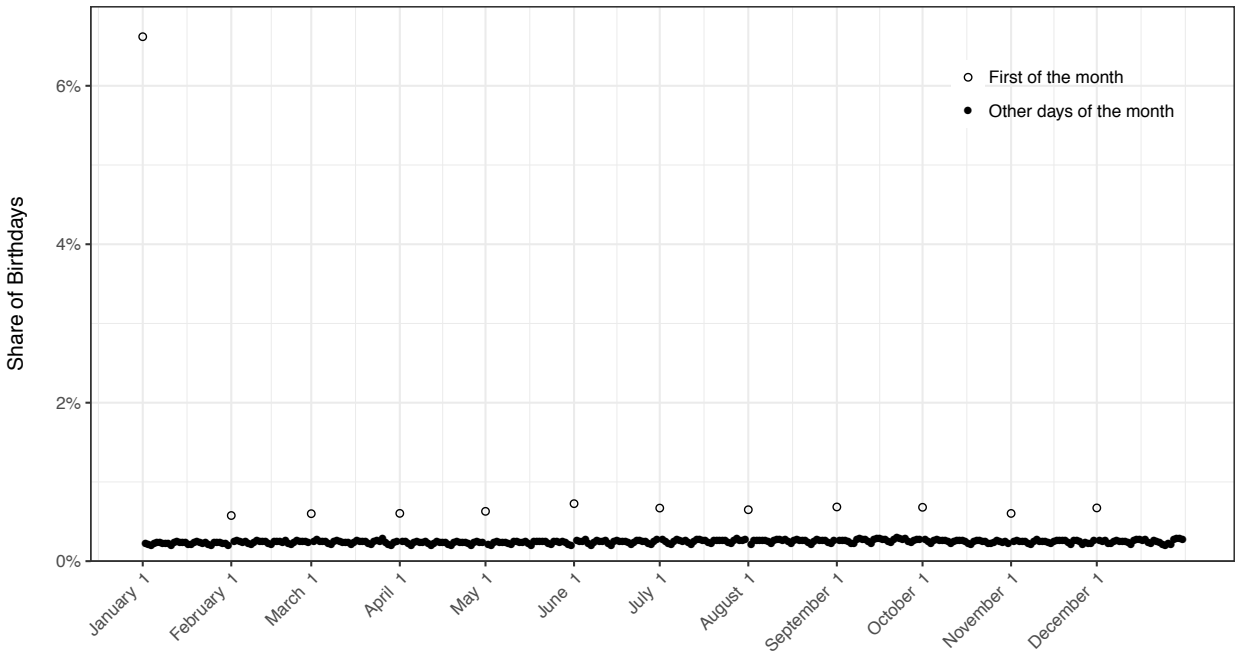


Figure A.2: Distribution of birthdays in 1970 in the voter file.

see that the District of Columbia, Arkansas, New Hampshire, Hawaii, and Wyoming also have a disproportionate number of cases in which voter records with the same observable characteristics reside in the same household. These issues in multi-generational households raise broader concerns about the quality of the voter file records in these states. We thus exclude these states from our preferred sample, and then scale-up our estimates to account for their removal when generating our final, national numbers.

Finally, we carry out a simulation to assess the sensitivity of our results to possible birthdate errors that may remain in our preferred sample. Given an error rate p, we randomly select $p\%$ of records in our preferred sample and assign each a new birthdate chosen uniformly at random from days in the recorded birth year. We then estimate the number of double votes in the synthetic dataset by running it through our full analysis pipeline, including estimation of $p_{b|f,l,y}$. Figure A.4 shows the result of this procedure when we simulate 10 synthetic datasets for each error rate p in the range 1% to 10%. We see that an error rate of p corresponds to an approximately $2p$ reduction in the estimated number of double votes.

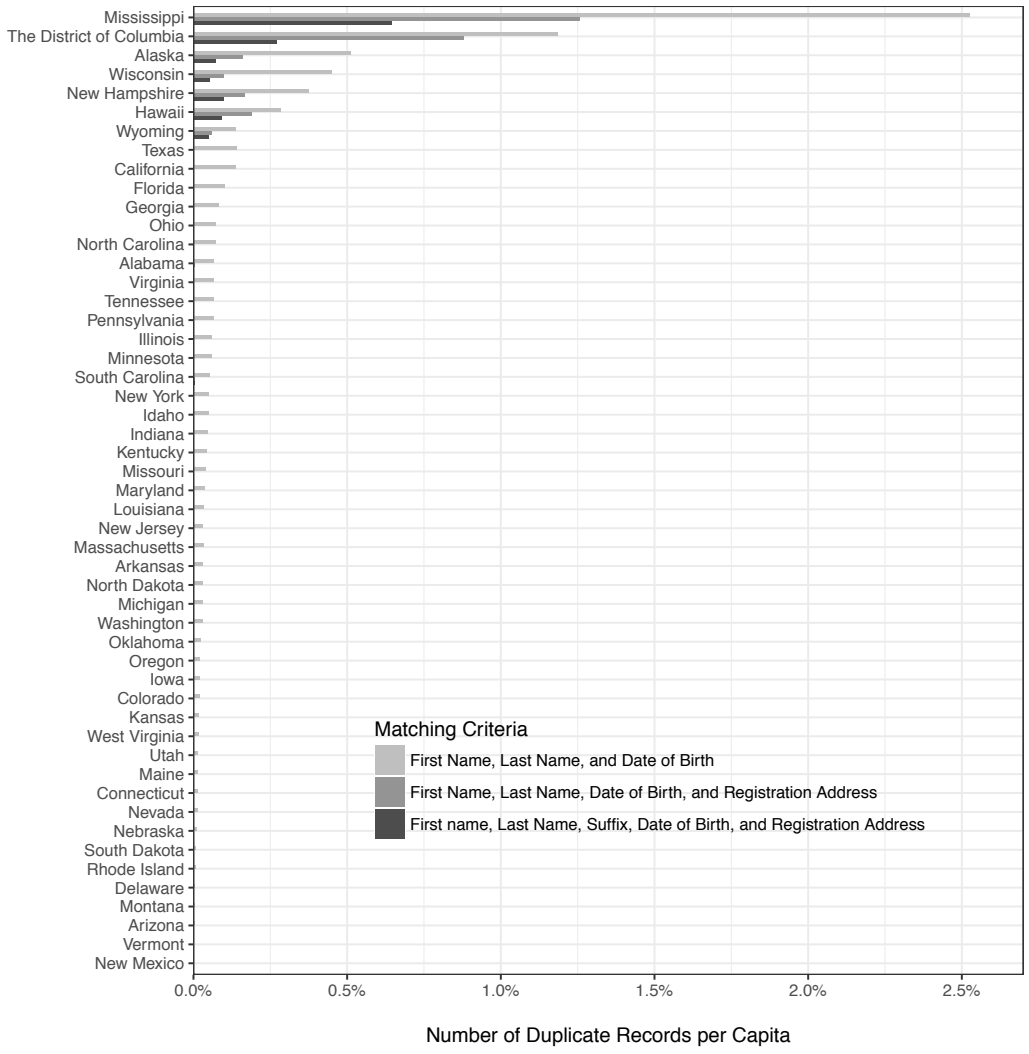


Figure A.3: Distribution of potential multi-generational matches within a state.

To understand why, note that any actual case of double voting in our synthetic datasets becomes undetectable with probability approximately equal to $2p$, since each vote record in the pair has probability p of being assigned a new birthdate. This explanation, however, only holds approximately, as birthdate errors also attenuate the day-of-week effect, among other factors, complicating theoretical analysis and prompting our simulation.

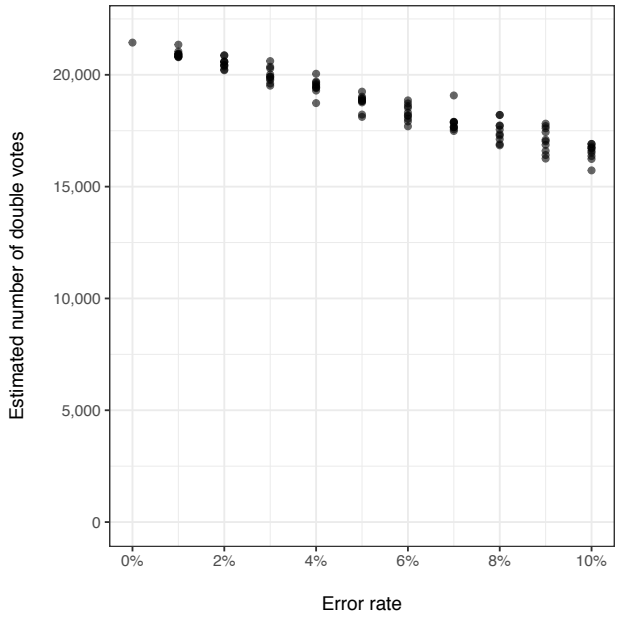


Figure A.4: Estimated number of double votes in the preferred sample from multiple simulations as we increase the error in recording of birthdates.

A.5 Evaluation on Synthetic Datasets

We evaluate the performance of our estimation strategy on synthetic datasets with a known number of double votes and which preserve key features of the real data, including correlations between names and dates of birth. To create each synthetic dataset, we carry out the following procedure, starting with the preferred version of the voter file.

1. Randomly select a year-of-birth and first name pair from the voter file.

2. Randomly, and independently of Step 1, select a last name from the voter file.

3. Given the selected first name, last name, and year of birth triple, generate a birthdate based on the modeled birthdate distribution $\hat{p}_{b|f,l,y}$.

4. Repeat the above three steps until the size of the sample equals the size of the voter file.

5. Randomly select k vote records in the synthetic dataset and add copies of them to the

synthetic dataset.

This procedure preserves the correlation between first names and dates of birth, including year. By randomly and independently selecting last names, we add additional variance to the dataset. Before duplicating any records, all observed matches are purely coincidental, and thus the full synthetic dataset has exactly k true double votes.

On each synthetic dataset, we carry out our full double vote estimation procedure, including fitting a model to estimate the distribution of $p_{b|f,l,y}$. Figure A.5 shows the result of this exercise on 100 synthetic datasets generated as above for a range of values for k. We find that our estimates are generally well aligned with the true number of double votes in these datasets. We also find that our analytic standard errors are, if anything, slightly too conservative. Specifically, among the 100 synthetic datasets, the analytic 95% confidence intervals always contained the correct value, and the 80% confidence intervals contained the correct value in 98 of the 100 instances.

We use an analogous simulation procedure to generate bootstrap estimates of variance for our empirical double vote estimate. Specifically, we generate 100 synthetic datasets as above, with k equal to our double vote point estimate, and then compute the variance of our 100 estimates on the synthetic datasets. This procedure can be viewed as a parametric bootstrap, as we use our estimated birthday model and point estimate of double votes to generate the bootstrap samples.

A.6 Estimating Errors in Recorded Voting

Ansolabehere and Hersh (2010) present the best evidence constructed to date on the accuracy of vote records in voter files. For each county in a given election, Ansolabehere and Hersh calculate the absolute value of the deviation between number of vote records in the voter file minus the total number of ballots cast in the certified aggregate returns. They aggregate these deviations over all of the counties in the state and divide by the total number

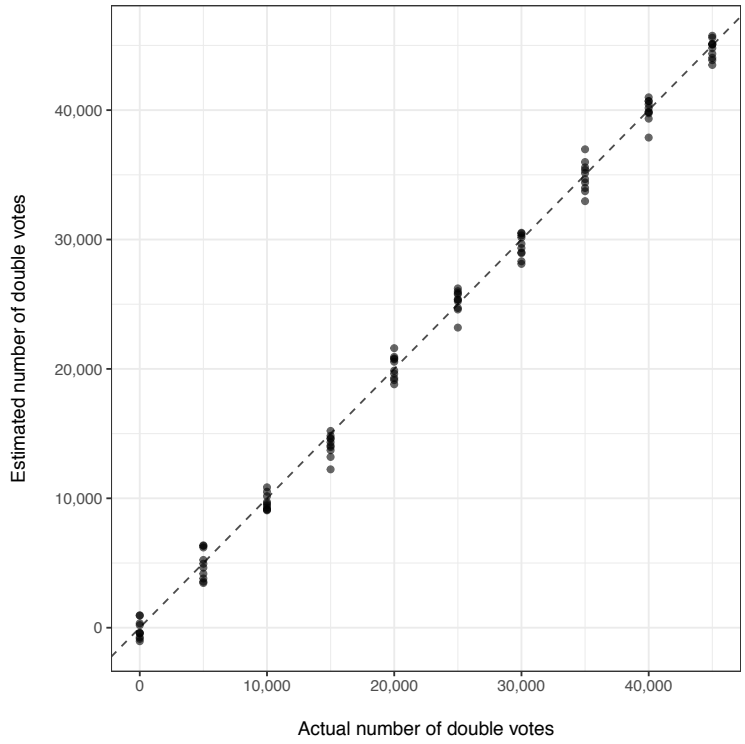


Figure A.5: Estimated number of duplicate records in a simulation compared to actual number of records duplicated.

of votes cast in the state. From this analysis, Ansolabehere and Hersh conclude that about two percent of voter registrations are incorrectly classified as having voted or abstained.

There are two primary limitations of this analysis. First, Ansolabehere and Hersh's method does not allow us to distinguish between false negatives and false positives, leaving open the possibility that there are few false positives. Second, their method also would understate the amount of measurement error in counties in which some registrations are wrongly classified as abstaining, while others are wrongly classified as voting.

We use the data collected from our Philadelphia poll book audit to estimate the rate at which registrations not used to vote are incorrectly given an electronic vote record (i.e., a false positive). There were 17,587 electronic registration records that did not have an electronic record of voting in these precincts.[22] In 33 of these cases, we found the registration had

[22] A few additional records could not be validated because pages were missing in the poll books.

a record of being used in the poll book. We also found 144 cases in which a registration was listed as voting in the electronic records, but had no record of having voted in the poll book (i.e., a signature discrepancy) and 29 cases of a registration being listed as voting in the electronic records, but not being listed in the poll book (i.e., a registration discrepancy). This suggests the false positive rate f_p is $\frac{144+29}{17{,}587+144+29-33} = 0.0098$.

Of course, we cannot be certain that these records are all false positives. It could be the case that the electronic voting records are correct and the poll book fails to note it. One way to indirectly assess this possibility is to compare the rates at which voter registrations with signature and registration discrepancies were recorded as voting in the elections leading up to 2010. If the previous vote history of these registrants is similar to the previous vote history of registrants who did not vote in 2010, this would suggest that many of these records are false positives. Conversely, if the previous vote history of these registrants is similar to the previous vote history of registrants who did vote in 2010, this would suggest that registrants with signature and registration discrepancies represent errors in the poll book, and thus are not false positives.

Table A.1 suggests that some, but not all, of the signature and registration discrepancies are false positives. To benchmark the past turnout of those who did and did not vote in 2010, we first calculate the 2006 turnout rate of those we know to have voted and not voted in 2010. Table A.1 shows that 62% of 2010 voters also turned out in 2006, while only 17% of those who abstained in 2010 participated in 2006. The 2006 turnout behavior of those with signature or registration discrepancies in 2010 falls somewhere in between, at 44% and 26%, respectively. We see similar patterns for 2007, 2008, and 2009 turnout as well. The fact that those with discrepancies between the electronic records and poll books previously voted at a rate somewhere in between those who abstained and those who voted in 2010 suggests that the false positive rate is both greater than zero and less than 1.0%.

These audit results are meant only to be illustrative, not representative, of the false positive rate in the population. There are some reasons why the false positive rate in

Table A.1: Examining Past Vote History of 2010 Signature and Registration Errors

	Dep. var.: Electronic record of general election voting in			
	2006	2007	2008	2009
	(1)	(2)	(3)	(4)
2010 electronic voting record	.448	.395	.436	.248
	(.005)	(.005)	(.005)	(.004)
Signature discrepancy	-.174	-.159	-.084	-.135
	(.042)	(.039)	(.033)	(.029)
Registration discrepancy	-.361	-.396	-.123	-.189
	(.079)	(.048)	(.076)	(.048)
Potential false negative	.224	.250	.357	.133
	(.085)	(.082)	(.067)	(.062)
Constant	.170	.083	.461	.018
	(.003)	(.002)	(.004)	(.001)

Note: N = 29,263 registered voters in the 47 precincts that were audited.

Philadelphia may be larger than the rate in the general population. Ansolabehere and Hersh (2010) found that there were more discrepancies than average in Pennsylvania between the number of ballots cast and the number of vote records in the voter file. And while a majority of jurisdictions either used Philadelphia's poll-book-and-bar-code approach or a voter sign-in sheet with no bar codes, a small, but growing number of jurisdictions, use an electronic poll book, particularly in states with early voting.[23] Because electronic poll books remove the step in which poll books are translated into electronic records, use of such technology is likely to reduce the number of false positives.

However, there are also reasons why we might expect there to be fewer false positives in Philadelphia than in the general population. Because of the size of the jurisdiction, the Philadelphia Voter Registration Office has a large, professionalized, and experienced staff that it can draw upon when scanning the poll books. And while there is more potential for error using the poll-book-and-bar-code approach than using electronic poll books, even more

[23]The Election Administration and Voting Survey suggests about 15% and 25% of voters used such technology in 2008 in 2012, respectively.

error is likely to occur in places that manually key-in the information contained in the poll book. It is also the case that there are false positives that our audit would not detect. For example, a poll worker could sign in a voter under the wrong registration. Consistent with this, Hopkins et al. (2017) report that 105 individuals had to resort to filing a provisional ballot in Virginia during the 2014 midterm election after they arrived at their polling place to find their registration was wrongly marked as having been used to vote earlier in the day.

Because we only have a rough sense of the rate of false positives, it is hard to say anything definitive about how many of the potential double votes can be explained by measurement error. Ultimately, all we can conclude is that measurement error likely explains a sizable portion, and possibly nearly all, of the surplus double votes that we observe in the national voter file.

A.7 Estimating the Number of Deadwood Registrations

As described above, the voter file incorrectly indicates some registrations were used to vote even though they were not, which can in turn affect estimates of double voting. To adjust for such errors, we need an estimate of the number of deadwood registrations for voters (c), as discussed in Section 4.3.

We follow a strategy similar to the one used in Theorem 1. While we cannot observe c directly, we can compute T, the number of observed cases in which two registration records in different states share the same first name, last name, and date of birth, and exactly one of them is recorded as having voted in the given election. As before, the estimator approximately subtracts from T the number of cases we would expect to observe due to chance in which a vote record and a non-voting registration record in different states share the same first name, last name, year of birth, and birthday given our estimates of $p_{b|f,l,y}$.

Our estimate of c involves four key assumptions that are analogous to our earlier ones. First, we assume that registration records are fully accurate. Second, we assume that each individual is at most registered in two states. Third, we assume that our estimate of the

birthday distribution, modeled as before, is accurate. Lastly, we assume individuals are listed in the poll books for a state if they have voted in that state in at least one of the two previous elections.

We start by decomposing c as the sum

$$c = \sum_f \sum_l \sum_y c_{f,l,y}, \tag{24}$$

where $c_{f,l,y}$ is the number of voters with first name f, last name l, and year of birth y who have a duplicate registration. Denote by $B_1, \ldots, B_q$ the birthdays for unique registration records with first name f, last name l, and birth year y. We assume these observed birthdays are $q \geq 1$ samples from a discrete probability distribution $D_{f,l,y}$ with values $b_1, \ldots, b_n$ and $\Pr_{D_{f,l,y}}(b) = p_{b|f,l,y}$. We further assume each of these registration records corresponds to one of u states we are analyzing named $\mathcal{S}_1, \ldots, \mathcal{S}_u$. We can enter cross-state duplicate registrations into our framework by assuming that there are k (with $0 \leq k \leq q$) duplicate records with birthdays $B_{q+1}, \ldots, B_{q+k}$, generated as $B_{q+i} = B_i$ and scattered in $\mathcal{S}_1, \ldots, \mathcal{S}_u$. Finally, we indicate whether observation B_i for $1 \leq i \leq q+k$ has been recorded as having voted or not by a flag f_i. In terms of this notation, $c_{f,l,y}$ is the number of duplicate pairs $\{(B_i, B_{q+i}) \mid 1 \leq i \leq k\}$ such that exactly one of the elements of the pair has voted, and $T_{f,l,y}$ is the number of pairwise matches among the $q+k$ observations such that the two elements of the pair are from different states and exactly one of them has voted. Theorem 2 below provides an estimator for $c_{f,l,y}$ based on $T_{f,l,y}$, $p_{b|f,l,y}$, and the number of recorded votes in each state.

Theorem 2. *Let v_l be the number of observations that voted in state $\mathcal{S}_l$ ($v_l = \sum_{B_i \in \mathcal{S}_l} f_i$), and $\bar{v}_l$ the number of observations without a vote in that state ($\bar{v}_l = \sum_{X_i \in \mathcal{S}_l} (1 - f_i)$). Define the estimator*

$$\hat{c}_{f,l,y} = \left(T_{f,l,y} - \left(\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l \right) \sum_i p_{b_i|f,l,y}^2 \right) \Bigg/ \left(1 - \sum_i p_{b_i|f,l,y}^2 \right). \tag{25}$$

Then $\mathbb{E}\hat{c}_{f,l,y} = c_{f,l,y}$ *and*

$$\mathrm{Var}(\hat{c}_{f,l,y}) \leq \left(\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l \right)^2 \left[\frac{\sum_i p_{b_i|f,l,y}^2}{1 - \sum_i p_{b_i|f,l,y}^2} \right]. \tag{26}$$

Proof. To simplify the notation, we represent $T_{f,l,y}$ by T, $D_{f,l,y}$ by D, $p_{b_s|f,l,y}$ by p_s, and $c_{f,l,y}$ by c. Let us first define $\mathcal{Q}$ to be the set of pairs (B_i, B_j) where $1 \leq i < j \leq q+k$, B_i and B_j belong to different states, and exactly one of them has its binary voting flag set to one. In other words

$$\mathcal{Q} = \{\ (B_i, B_j)\ \mid\ 1 \leq i < j \leq q+k\ \ ,\ \ 1 \leq \nexists u \leq l : \{B_i, B_j\} \subset \mathcal{S}_u\ \ ,\ \ f_i \oplus f_j = 1\ \}.$$

Here, $f_i \oplus f_j = 1$ means exactly one of f_i and f_j is set to one.

Based on this notation, T is the number of pairs $(B_i, B_j) \in \mathcal{Q}$ such that $B_i = B_j$, and c is the number of cases for $1 \leq i \leq k$ where $(B_i, B_{q+i}) \in \mathcal{Q}$.

Let $A_{i,j}$ indicate whether $B_i = B_j$. Then by the linearity of expectation,

$$\mathbb{E}T = \mathbb{E}\left(\sum_{(B_i,B_j)\in\mathcal{Q}} A_{i,j} \right) = \sum_{(B_i,B_j)\in\mathcal{Q}} \mathbb{E}A_{i,j}. \tag{27}$$

For all the (B_i, B_j) pairs in $\mathcal{Q}$ for which $j = q+i$, $B_i = B_j$ by construction, so $\mathbb{E}A_{i,j} = 1$. By definition, the number of these pairs is c. For the remaining $|\mathcal{Q}| - c$ pairs, $\mathbb{E}A_{i,j} = \Pr_D(B_i = B_j) = \sum_s p_s^2$. Consequently,

$$\begin{aligned} \mathbb{E}T &= c + (|\mathcal{Q}| - c) \sum_s p_s^2 \\ &= c\left(1 - \sum_s p_s^2\right) + |\mathcal{Q}| \sum_s p_s^2. \end{aligned}$$

To compute $|\mathcal{Q}|$, we first count all the (B_i, B_j) pairs where $i < j$ and exactly one of f_i and f_j is set to one. This count is equal to number of ways we can choose a pair with first element from

observations with flag set to one ($\sum_{l=1}^{u} v_i$ observations) and second element from observations with flag set to zero ($\sum_{l=1}^{u} \bar{v}_i$ observations), which sums up to $\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l$. Then we eliminate the pairs where B_i and B_j are from the same set. For each set $\mathcal{S}_l$, we need to eliminate $v_l \bar{v}_l$ such pairs. Therefore,

$$|\mathcal{Q}| = \sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l.$$

By substituting $|\mathcal{Q}|$ and rearranging terms, we now have that $\mathbb{E}\hat{c} = c$.

To compute the variance of $\hat{c}$, we first decompose variance of T as

$$\mathrm{Var}(T) = \sum_{(B_i,B_j)\in\mathcal{Q}} \mathrm{Var}(A_{i,j}) + 2\sum_{\mathcal{R}} \mathrm{Cov}(A_{i,j}, A_{k,l}) \tag{28}$$

where $\mathcal{R}$ is the set of (i, j, k, l) indices such that each distinct unordered pair from elements in $\mathcal{Q}$ appears in the sum exactly once. For $A_{i,j}$ we can write,

$$\mathrm{Var}(A_{i,j}) = \mathbb{E}A_{i,j} - (\mathbb{E}A_{i,j})^2 . \tag{29}$$

For all the (B_i, B_j) pairs in $\mathcal{Q}$ for which $j = q + i$, $\mathbb{E}A_{i,j} = 1$. Therefore, for those pairs $\mathrm{Var}(A_{i,j}) = 0$. There are c such pairs in $\mathcal{Q}$, and for the remaining $|\mathcal{Q}| - c$ pairs, $\mathrm{Var}(A_{i,j}) = \sum_s p_s^2 - (\sum_s p_s^2)^2$. Consequently,

$$\sum_{(B_i,B_j)\in\mathcal{Q}} \mathrm{Var}(A_{i,j}) = (|\mathcal{Q}| - c)\left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{30}$$

Next we consider the covariance terms $\mathrm{Cov}(A_{i,j}, A_{k,l})$. By Cauchy-Schwarz's inequality,

$$\mathrm{Cov}(A_{i,j}, A_{k,l}) \leq \sqrt{\mathrm{Var}(A_{i,j})\mathrm{Var}(A_{k,l})}. \tag{31}$$

If either (B_i, B_j) or (B_k, B_l) are among the c pairs in $\mathcal{Q}$ for which one observation is a

copy of another, then $\text{Var}(A_{i,j})\text{Var}(A_{k,l}) = 0$. For all the other cases, $\text{Var}(A_{i,j})\text{Var}(A_{k,l}) = \left(\sum_s p_s^2 - (\sum_s p_s^2)^2\right)^2$. Therefore,

$$\sum_{\mathcal{R}} \text{Cov}(A_{i,j}, A_{k,l}) \leq \binom{|\mathcal{Q}| - c}{2} \left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{32}$$

Combining equations for terms in $\text{Var}(T)$, we can write,

$$\text{Var}(T) \leq (|\mathcal{Q}| - c)^2 \left(\sum_s p_s^2 - \left(\sum_s p_s^2\right)^2\right). \tag{33}$$

Consequently,

$$\begin{aligned} \text{Var}(\hat{c}) &= \text{Var}(T) / \left(1 - \sum_s p_s^2\right)^2 \\ &\leq (|\mathcal{Q}| - c)^2 \left[\frac{\sum_s p_s^2}{1 - \sum_s p_s^2}\right]. \end{aligned}$$

To make the bound on $\text{Var}(\hat{c})$ independent of c, we substitute $|\mathcal{Q}| - c$ by $|\mathcal{Q}|$ and replace it with the previously calculated count, which yields to

$$\text{Var}(\hat{c}) \leq \left(\sum_{l=1}^{u} v_l \sum_{l=1}^{u} \bar{v}_l - \sum_{l=1}^{u} v_l \bar{v}_l\right)^2 \left[\frac{\sum_s p_s^2}{1 - \sum_s p_s^2}\right].$$

□

A.8 Measurement Error Linking Vote Records to Crosscheck Data

We calculated the frequency with which votes are cast using the registration records flagged by Crosscheck by merging the Crosscheck data with the Target Smart national voter file by exactly matching records in the two data sources on first name, middle name, last name, date of birth, and state. Doing so potentially could cause us to under- or overestimate the rate at which the registrations flagged by Crosscheck were used to vote. We would

underestimate turnout if records for the same person did not exactly match on one of these five variables. Conversely, we would overestimate turnout if a registration identified by Crosscheck matched with a district person's vote record in the Target Smart data. To get a sense of which, if either, of these sources of measurement error are a bigger issue, we take advantage of the fact that we know a registrant's voter registration number if they are registered to vote in Iowa. Thus, we compare the vote history we estimate when we match to Target Smart to the vote history we estimate when we directly link the Crosscheck data to the Iowa voter file using the voter registration number.

Table A.2 suggests that measurement error in turnout does not affect our conclusion that few likely double votes were identified in the Crosscheck data. Columns 3 and 4 replicate our 2012 analysis when Iowa turnout is linked to the Crosscheck data from the voter file using Iowa's voter registration number. While we find one additional case of a likely double vote, we also find more than a hundred additional cases in which only the Iowa registration was used to cast a vote. We expand upon this analysis in columns 5 and 6 by limiting the sample of states paired to Iowa to those states in which fewer than 10% of 2012 voters have a birthday on the first of the month. We do this because we expect there to be fewer cases in which we fail to match a vote record to a registration record in these states. We find that 7 of the 1,076 potential double votes were actually double votes in these states. Moreover, we find 1,994 cases in which only the voter registration record with the earlier registration date was used to cast a ballot.

Table A.2: Robustness Checks on 2012 Analysis in Table 1

Target Smart (TS) or Vote File (VF) to Measure Iowa Turnout	TS		VF		VF	
Drop States with > 10% First of Month Birthdays	No		No		Yes	
SSN4 Match	Yes	No	Yes	No	Yes	No
Which Reg. Used to Vote:						
Both	7	1476	8	1489	7	1069
One (earlier reg. date)	2542	1678	2694	1748	1994	1117
One (later or unknown reg. date)	9430	2581	9883	2657	7843	2225
Neither	14008	3178	13402	3019	8817	2085

A.9 Additional Tables and Figures

Table A.3: Estimated Double Votes in Sample by Sample Restriction and Birthday Distribution

	(1)	(2)	(3)	(4)
Drop First of Month Birthdays	No	Yes	Yes	Yes
Drop States with Multigenerational Issues	No	No	Yes	Yes
Keep Commercially Sourced Birthdays	Yes	Yes	Yes	No
Number of Vote Records in Sample (millions)	124.94	107.62	104.21	102.65
Coverage in Sample of Total Votes (FEC)	.968	.834	.807	.795
Vote Record Pairings with Same First and Last Name and DOB	3050762	827840	763133	738670
Estimated Double Votes in Sample by Distribution:				
Birthday Distribution Uniform	2067049	47523	28921	27886
	(1992)	(1769)	(1716)	(1689)
Birthday Distribution Conditional on Year	154687	44335	26080	25137
	(7797)	(1774)	(1720)	(1693)
Birthday Distribution Conditional on Year and First Name	410323	39723	21724	20887
	(6917)	(1782)	(1728)	(1700)

Table A.4: Robustness Checks on Estimated Double Votes in Population

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Drop States with Multigenerational Issues	Yes	No	Yes	Yes	Yes	Yes	Yes
Standardize First Names	Yes	Yes	No	Yes	Yes	Yes	Yes
Smoothing Parameter in Birthday Distribution (θ)	11000	11000	11000	1000	50000	11000	11000
Keep Observations With Commercial Birthday	Yes	Yes	Yes	Yes	Yes	No	Yes
Analytical (A) or Bootstrapped (B) Standard Errors	A	A	A	A	A	A	B
Estimated Double Votes in Sample	21724	39723	20525	20103	23270	20887	21724
Sample-to-Population Scale Factor	1.534	1.439	1.534	1.534	1.534	1.580	1.534
Estimated Double Votes in Population	33346	57161	31506	30858	35719	33022	33346
	(2652)	(2564)	(2465)	(2656)	(2649)	(2688)	(1302)[a]

[a]The 95% bootstrap percentile confidence interval is [30,037, 35,373].

	Grid of Potential Duplicate Voters Within States by DOB Last Name First Name														
2012	AZ	AR	CO	IL	IA	KS	KY	LA	MI	MS	MO	NE	OK	SD	TN
AZ		2,829	24,863	16,014	7,153	3,687	688	2,062	27,617	2,220	7,569	3,306	4,006	2,449	3,614
AR	2,829		4,557	6,950	2,430	2,686	691	5,957	5,085	6,477	11,049	995	7,403	433	7,180
CO	24,863	4,557		19,902	10,850	10,035	1,054	5,065	17,086	3,309	12,498	8,927	8,306	3,937	6,153
IL	16,014	6,950	19,902		31,882	6,311	2,467	5,207	49,260	10,766	39,658	3,803	4,834	1,500	12,469
IA	7,153	2,430	10,850	31,882		4,706	526	1,558	7,019	1,797	11,563	10,954	2,031	4,865	2,806
KS	3,687	2,686	10,035	6,311	4,706		401	1,369	4,461	1,397	31,082	4,196	6,575	905	2,205
KY	688	691	1,054	2,467	526	401		873	2,267	1,085	1,195	233	576	117	1,905
LA	2,062	5,957	5,065	5,207	1,558	1,369	873		6,851	17,744	5,254	810	2,829	277	4,422
MI	27,617	5,085	17,086	49,260	7,019	4,461	2,267	6,851		7,527	12,960	2,416	4,067	1,265	16,956
MS	2,220	6,477	3,309	10,766	1,797	1,397	1,085	17,744	7,527		5,607	780	2,364	305	21,661
MO	7,569	11,049	12,498	39,658	11,563	31,082	1,195	5,254	12,960	5,607		4,244	7,539	1,300	7,804
NE	3,306	995	8,927	3,803	10,954	4,196	233	810	2,416	780	4,244		1,126	2,608	1,108
OK	4,006	7,403	8,306	4,834	2,031	6,575	576	2,829	4,067	2,364	7,539	1,126		402	2,858
SD	2,449	433	3,937	1,500	4,865	905	117	277	1,265	305	1,300	2,608	402		537
TN	3,614	7,180	6,153	12,469	2,806	2,205	1,905	4,422	16,956	21,661	7,804	1,108	2,858	537	
Totals	108,077	64,722	136,542	211,023	100,140	80,016	14,078	60,278	164,837	83,039	159,322	45,506	54,916	20,900	91,678

Figure A.6: Distribution of potential duplicate voters in 2012 according to internal documents circulated by the Interstate Crosscheck Program.

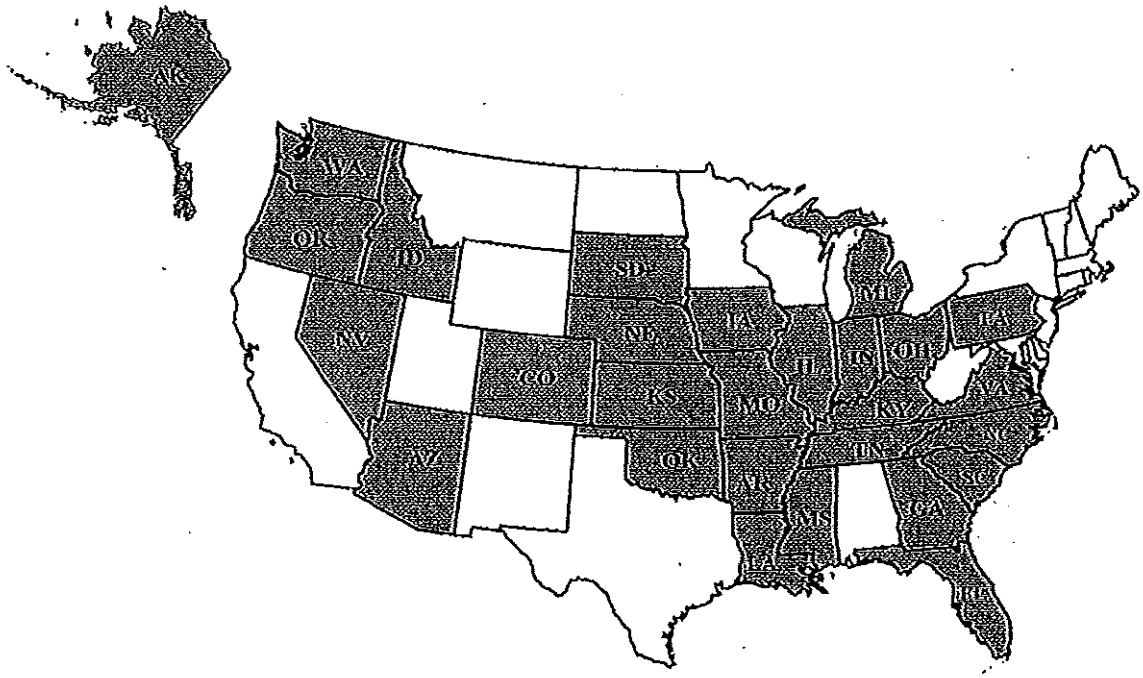
A.10 Crosscheck 2014 Participation Guide

Interstate Voter Registration Data Crosscheck

2014 Participation Guide

December, 2013

Alaska, Arizona, Arkansas, Colorado, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Michigan, Mississippi, Missouri, Nebraska, Nevada, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, South Dakota, Tennessee, Virginia and Washington.



Contents

I. Joining the Crosscheck Program

1. Chief State Election Official (CSEO) or designee signs the Memorandum of Understanding (MOU)

2. CSEO assigns two staff members:
    a. one election administration person
    b. one IT/database person

3. Staff members will:
    a. participate in annual conference calls and emails
    b. pull voter registration data in January and upload to FTP site
    c. receive crosscheck results and process them
    d. instruct local election officials
        (1) mail notices to registrants
        (2) promptly respond to requests for signatures, addresses, etc.

4. There is no cost. Processing the duplicate registrations and researching possible double votes requires a commitment of time at the state and local levels. States make individual decisions about the amount of time and effort they will commit, and this might vary from year to year. A state that is not able to commit the resources to process the results in a given year still provides a benefit to the other states through its participation.

II. Data Comparison Procedure

1. Designate at least one administrative and one IT/database contact person to be on the email list serve.

2. At least one person from each state should participate in a conference call hosted by Kansas in November or December preceding the crosscheck.

3. Pull your entire database on January 15, 2014 and upload it to the secure FTP site hosted by Arkansas. Instructions and, if necessary, followup reminders and questions, will come from the Kansas or Arkansas IT/database persons. Each state's data should include Active and Inactive records if possible.
Follow the prescribed data format. Review and edit your data before uploading it. Do not truncate fields, reverse fields, or leave them blank. Please include SSN4.

4. Kansas will download files, delete everything from FTP site, run the data comparison, and upload individual state results files to the FTP site. At every stage of the process, data files are encrypted and zipped.

5. When notified by Kansas, download your results files from the Arkansas FTP site. After downloading, make sure all data are deleted from the FTP site.

6. Process the results according to your state's laws, regulations and policies.

7. Respond promptly to inquiries from other states or local jurisdictions for information to confirm duplicates or to obtain evidence of double votes. Usually this will be copies of signatures on poll books or absentee/advance ballot applications and return envelopes.

III. Analyzing Results

Each state analyzes and acts upon the results according to its own laws and regulations. No state is required by the Memorandum of Understanding to act upon the results.

A. Cancellations and Confirmation Mailings

An apparent duplicate registration is produced when the first names, last names and dates of birth in two records match exactly. Other information such as middle name, suffix and SSN4 should be used to confirm whether the two records are matches. It may be necessary to contact another jurisdiction to obtain more information, such as signatures.

An apparent duplicate registration may result in one of two actions being taken:

1. The jurisdiction possessing the record with the older registration date may cancel the record (and send a cancellation notice if state laws or regulations require it) if the following conditions are met:
    a. The records match on first name, last name, and date of birth, and
    b. One or both of the following data elements match:
        - last four digits of Social Security number and
        - signature and
    c. Data in the middle name field either matches or is not a mismatch.

2. The jurisdiction possessing the record with the older registration date may mail a confirmation notice, pursuant to the National Voter Registration Act of 1993, Sec. 8(d)(2), if the three fields match as specified in item 1.a. above. These registrants' names are added to the state's Inactive list pending cancellation after two federal general elections, assuming there has been no voting activity during that period.

B. Cancellations by Confirmation Between Jurisdictions

Pursuant to NVRA Sec. 8(d)(1)(A), the jurisdiction possessing the record with the older registration data may cancel the record (and send a cancellation notice if state laws or regulations require it) if another jurisdiction confirms that the registrant has registered to vote in the newer jurisdiction and has indicated on the voter registration application form an address in the former jurisdiction.

C. Information Sharing

Each state will decide whether it prefers that followup requests for information from other states and localities should be addressed to the state or the individual

localities. All participating states will be notified of this preference. Each state will provide contact information for local election offices. States and localities are cautioned against sending registrants' personally identifiable information via email.

D. Double Votes

1. When two records are determined to be duplicates, review the voter history field to determine if there appears to be a double vote. Experience in the crosscheck program indicates that a significant number of apparent double votes are false positives and not double votes. Many are the result of errors—voters sign the wrong line in the poll book, election clerks scan the wrong line with a barcode scanner, or there is confusion over father/son voters (Sr. and Jr.).

2. Collect copies of signatures from the election officers in the two jurisdictions in which the double votes occurred. The classic double vote occurs when a person votes in person at the polling place on election day in the jurisdiction where he/she normally lives and also casts an absentee (advance) ballot by mail in the other jurisdiction.
In these cases, evidence to prove the double vote occurred often includes the following:

- Signature from the voter's application for voter registration in jurisdiction A
- Signature from the voter's application for voter registration in jurisdiction B
- Signature from the poll book in jurisdiction A
- Signature on an absentee (advance) ballot application form in jurisdiction B
- Signature on the absentee (advance) ballot return envelope in jurisdiction B

3. The collection of evidence to prove double votes is a considerable commitment of time and effort. It requires a high level of cooperation and communication between jurisdictions.

4. Compare the signatures. Once you are satisfied that the evidence indicates a double vote occurred, refer the case to a local or state prosecutor. Include a referral cover letter, cite relevant state statutes, and include copies of all necessary documents.

E. Information Request Form

A request form as appears below may be used to request followup information from other jurisdictions. Jurisdictions may adapt it as needed and produce it on their own letterhead.

KRIS W. KOBACH
Secretary of State

Memorial Hall, 1st Floor
120 S.W. 10th Avenue
Topeka, KS 66612-1594
(785) 296-4564

STATE OF KANSAS

12/4/2013

To Whom It May Concern:

Voter Information

John Doe	DOB:	1/02/1933
Jane Doe		2/03/1955
Tom Smith		3/06/1985
Janet Jones		12/13/1967
Ben Thompson		11/9/1990

We request voter registration and voter history information related to the above mentioned individuals for the November 2012 election. The purpose of the request is to collect evidence about possible double votes cast by these individuals.

We will maintain appropriate safeguards to protect the confidentiality of the records.

We will not make any public use of these files or information. We will keep your office apprised of the details as our office moves forward with this inquiry.

If you have questions please contact me at 785-296-0080.

Sincerely,

Jameson Beckner
Special Programs Coordinator
Kansas Secretary of State

IV. IT/Database Manager Information

IT/database managers should follow this timeline and use the data format on the next page.

ACTIVITY	Time Frame
Kansas sends data extract reminder email	January
Arkansas sends upload instruction email to each state with: • URL for FTP site • login ID • password	January
States extract their data according to Data Format document	approx. January 15th
States upload their extract files to the FTP site	January
Each state emails bruce.ferguson@sos.ks.gov • with encryption password • with number of records	January
Kansas processes the extract file	January - February
Kansas emails notification to each state	January - February
Kansas loads the file into comparison database	January - February
Kansas produces Results file for each state: • create Single Row comparison files • create Stacked Row comparison files • update Statistics spreadsheet • zip all comparison files and statistics into Results file • encrypt Results file into self-decrypting .exe • upload Results file to the FTP site	January - February
Kansas sends email to each state that Results file is ready	January - February
Each state needs to: • refer to State Cross Check Result File Instructions below • download their Results file • delete their Results file from the FTP site • decrypt and unzip their Results file	January - February
Each state processes its Results file accordingly	January - February
Kansas and Arkansas verify Results have been deleted from FTP site	February

Data Format

Fields	
1. Status	("A" – Active, "I" – Inactive)
2. DateTime_Generated	
3. First_Name	
4. Middle_Name	
5. Last_Name	
6. Suffix_Name	
7. Date_of_Birth	(YYYY/MM/DD Example: "2010/01/01")
8. Voter_ID_Number	
9. SSN_Last4	
10. Address_Line_1	(if no mailing address, provide residential address)
11. Address_Line_2	
12. City	
13. State	
14. Zip	
15. County_Name	
16. Date_of_Registration	(YYYY/MM/DD Example: "1970/01/01")
17. Voted_in_Last_General	("Y" – they did vote, or "N" – they did not vote, or "" – data not available)

The file should be a comma delimited ASCII file with double quote text qualifiers and {CR} {LF} row delimiters. The file should have a Header Record followed by 1 to many Voter Records. Each Voter Record should contain 17 fields.

Example:

"Status","DateTime_Generated","First_Name","Middle_Name","Last_Name","Suffix_Name", "Date_of_Birth","Voter_ID_Number","SSN_Last4","Address_Line_1","Address_Line_2", "City","State","Zip","County_Name","Date_of_Registration","Voted_in_Last_General"

"A","2013/01/15 12:00:00 AM","Bob","Alan","Jones","","1940/06/16","123456","7890", "123 Main St","Apt 201","Topeka","KS","12345","Shawnee","1958/06/17","Y"

The file should be encrypted and password protected and uploaded to the secure FTP site. Please email the password in a separate email.
Also, please notify us of the total number of records in the uploaded file.

We use a free program, AxCrypt, for encryption. Here is a link to the AxCrypt download site: http://www.axantum.com/AxCrypt/ .

Reminders for Data Upload Process

We have identified from past experience some helpful hints that we ask you to keep in mind as you prepare to upload your data. Please carefully review your file before uploading, taking into account the following:

1. Do not include any records that contain programming commands from your process that created the file
2. Please include a header record, but only one
3. Please account for all 17 fields, in the order requested
4. Please trim all excess spaces so the records are not padded to a fixed length
5. Please zip your file before uploading it to the FTP site
6. Please be aware if your address lines contain a comma and make sure you encapsulate the field with double quotes
7. Please note that if you will be providing SSN data, we only ask for the last 4 positions
8. Please extract dates in the requested formats
9. Please edit data that contains double quotes during extraction – ie. remove the double quotes or change them to single quotes
    a. Examples
        1.. Change …,"Robert "Bob"",…
        To …,"Robert 'Bob"",…

        2. Change …,"123 "U" St",…
        To …,"123 U St",…
10. Consider replacing "null" text values with actual null string, ""

Crosscheck Results File Instructions

1) Download your state's self decrypting .exe file from the FTP site
2) Double click the file.
3) Enter the passphrase
    i) This will decrypt the file
    ii) The resulting .zip file contains 2 folders and a spreadsheet
4) Extract all files
5) Please delete your state's file from the FTP site once you confirm a successful download

- The spreadsheet presents some general statistics about current and previous State Cross Check Voter Registration Comparisons
  - Since DOB is one of the match criteria, please provide valid Date_of_Birth fields (see column D)
  - If you see a non-zero value in column E, please determine if you can provide unique Voter_ID_Number fields
  - If you see a non-zero value in column F, please determine within your own extract file if you have multiple records for the same individual (the criteria for this comparison is the same as the state-to-state comparison; DOB, LastName, FirstName

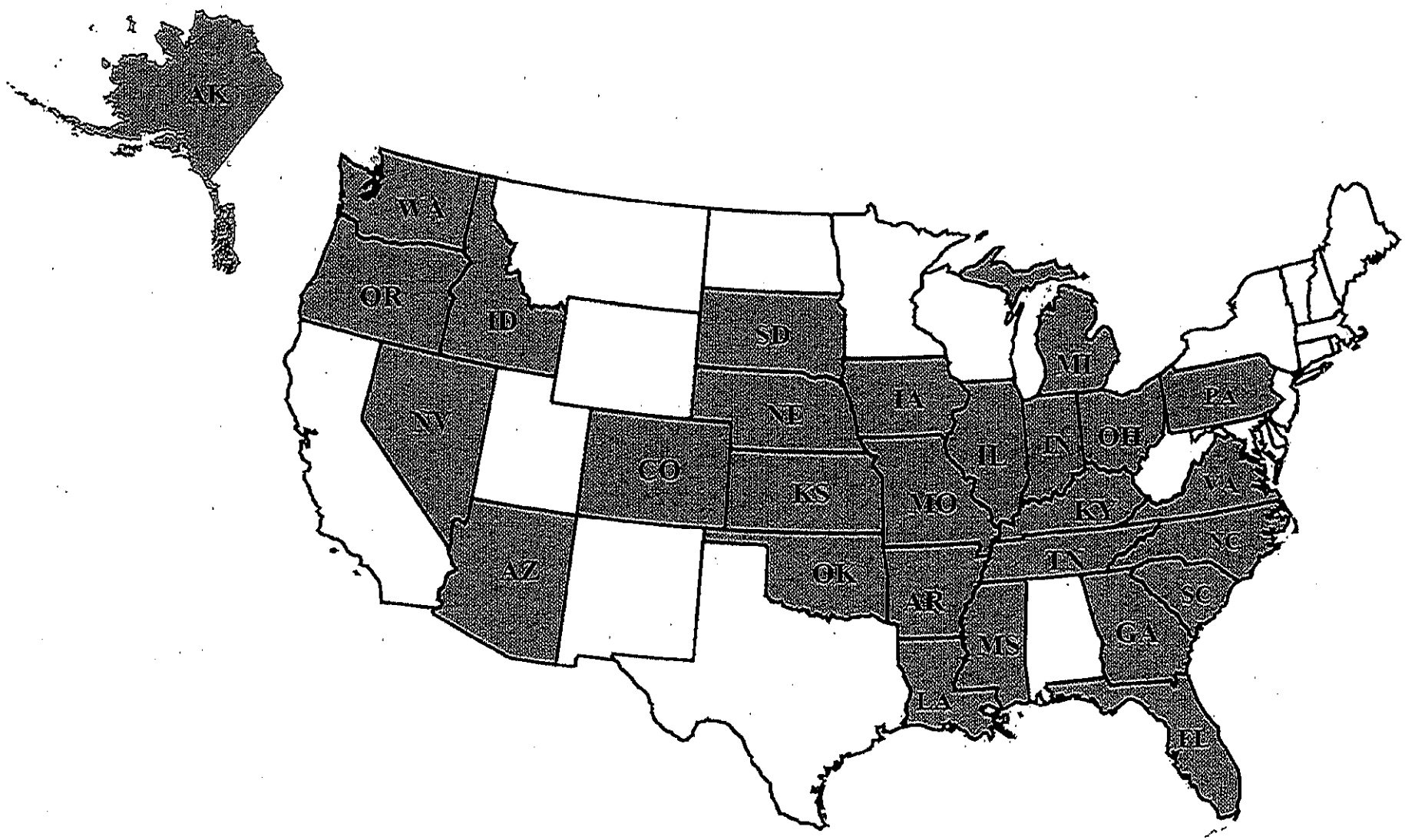
- One folder, SingleRowOutput:
  - Contains one result file with the potential match count of each comparison and total for that BaseState
  - Contains individual result files for your state compared with each other participating state
    - Ideally opened programmatically
  - Contains one result file comparing your state with all other participating states
    - Ideally opened programmatically
  - Within each result file:
    - A header row identifies each column
    - A possible voter match is presented in a single row with your states' data followed by the data from the other state

- The second folder, StackedRowOutput:
  - Contains one result file with the potential match count of each comparison and total for that BaseState
  - Contains individual result files for your state compared with each other participating state
    - Ideally opened in Excel
  - Contains one result file comparing your state with all other participating states
    - Ideally opened in Excel
  - Within each result file:

- A header row identifies each column
- A possible voter match is presented in two rows with data from your state stacked over data from the other state

➢ Within each individual result file:
- The "Case" column represents the sequential instance of each possible match

➢ Within the ALL result file:
- The "Case" column represents the sequential instance of each possible match as that match relates in the individual result files



2014 Interstate Crosscheck