RLA Overview

Draft: January 9, 2019

Kellie Ottoboni

**Math and Tools**

These pilots were the first time the SUITE method for hybrid RLAs has been used in practice. SUITE is a general method for conducting RLAs of *stratified samples* of ballots, where ballots are divided into distinct non-overlapping groups (or *strata*)and samples of ballots are drawn independently from each group. Audits can essentially be conducted separately, but in parallel, for each stratum of ballots and then combined to produce a single risk limit. In particular, SUITE is useful for auditing contests that include jurisdictions with heterogeneous voting equipment. Ballot polling can be conducted alongside ballot-level comparison audits of ballots cast on equipment that produces ballot-level CVRs. The math requires minor modifications to existing ballot polling and ballot-level comparison RLAs. By using just one stratum, the SUITE method reverts to the single RLA method of choice. (In Rochester Hills, we used SUITE ballot polling for the entire contest.)

Kellie Ottoboni and Philip Stark created a software tool for SUITE. The code is written in Python and displayed in an interactive Jupyter Notebook. The tool can be run interactively at https://mybinder.org/v2/gh/pbstark/CORLA18/master?filepath=code\%2Fsuite\_toolkit.ipynb.

This was intended to be a prototype user interface for these pilot RLAs rather than an industrial strength tool; the hope is that software developers will be able to base a proper program on our prototype.

The SUITE tool calculates the necessary pieces of information for each step of the audit:

* It estimates the initial sample sizes needed in each stratum to stop the audit, assuming that the reported results were correct and that one-vote overstatements occur at a very low rate in the ballot comparison stratum.
* It uses the SHA256 PRNG from the [cryptorandom package](http://www.github.com/statlab/cryptorandom) to draw the random samples of ballots in each stratum.
* It reads in a ballot manifest for each stratum as separate CSV files, then determines which ballots in which batch need to be pulled based on the samples that were drawn previously, and finally exports this data to a CSV file to be printed.
* It takes the statistics from each sample (the number of 1-vote and 2-vote overstatements and understatements in the comparison stratum, and the number of votes seen for each candidate in the polling stratum) as input and runs the risk calculation for each pair of reported winners and reported losers.
* If not all winner, loser pairs are confirmed, it estimates how many more ballots need to be sampled in each stratum to confirm the reported outcome, assuming that the rates of discrepancies and the rates of votes for each candidate will continue to reflect those seen in the initial samples.
* It runs each of these steps again for a second round of sampling.
* It logs each step of the process in JSON files.

**Procedures**

In coordination with the Bureau of Elections at the Michigan Secretary of State's office,

we performed three pilots of risk-limiting audit procedures. This section describes the audits in each city and the differences between them. The table below summarizes the sizes of each audit.

Since the goal of these pilots was to gain hands-on experience, instruct local election officials, compare procedures, and identify bottlenecks in the processes, we opted to reduce the population of ballots under audit to each city. The sampling frame for a true RLA includes all ballots cast for a particular contest; none of the contests under audit were entirely contained in the sampling frame of ballots. Each audit pretended that the ballots we had access to comprised the entire contest. Nonetheless, reducing the sampling frame made these pilots feasible to conduct in a day and illustrated RLA procedures to a wide audience.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| City | Total ballots | Winner | Loser | Margin | % Ballots with a CVR | # ballots audited | % ballots audited |
| Rochester Hills | 36,666 | 22,999 | 12,343 | 29% | 0% | 76 | 0.2% |
| Lansing | 21,328 | 10,309 | 7,694 | 12% | 50% | 260 | 1.2% |
| Kalamazoo | 27,666 | 20,699 | 5,569 | 55% | 19% | 40 | 1.4% |

**Rochester Hills**

The Rochester Hills pilot RLA took place on December 3. About 30 local election officials from around the state attended to observe the audit process. We audited Proposal 18-3, a statewide proposal to add new voting policies to the Michigan constitution. In Rochester Hills, “Yes” votes won with a 29% margin.

Rochester Hills uses Hart InterCivic Verity tabulators, which do not provide ballot-level CVRs. Thus, we used entirely ballot polling for the audit. We did not need to use different software for this: The SUITE tool reverts to pure ballot polling when the CVR stratum size is 0 ballots.

The SUITE tool estimated that we'd need to look at 76 ballots to achieve a risk limit of 5% if the reported results were correct. We had 5 two-person audit boards pull the ballots from ballot bags and sample them either by counting down to the ballot that the SUITE tool sampled or by using $k$-cut to select a ballot from the stack. Audit boards used a mix of counting and $k$-cut to sample ballots: if the SUITE tool sampled a ballot that would require counting more than 200 ballots to find it, the team used $k$-cut instead.

It took about 2.5 hours for the audit boards to pull the ballots and record them, and another half hour for a separate team to tally them.

The sample contained 50 “yes” votes and 26 “no” votes. The risk of the sample was 2%.

**Lansing**

The Lansing pilot RLA took place on December 4. We audited the 54-A District Court Judge race using a 9% risk limit. Cynthia Ward won over Ayanna Neal by about a 10% margin. Though this contest was contained within the city of Lansing, we reduced the sampling frame of ballots for other reasons, making this a pilot of RLA procedures rather than a true RLA.

First, there were several ballot bags that could not be opened. Precinct 45 was one of the randomly selected precincts to do a full hand recount for the post-election audit procedures developed under MCL 168.319. As such, the ballot bags containing ballots from Precinct 45 could not be opened. These bags contained ballots from other precincts as well. One way to handle this would have been the “phantoms to zombies” approach,

treating each inaccessible ballot in the sample as a vote for the reported loser (Banuelos and Stark, 2012). Because the point of the pilot was to illustrate procedures rather than to obtain a precise risk measurement, we opted to remove ballots from the ballot bags containing Precinct 45 from the sampling frame.

Second, the SUITE tool estimated that the initial sample size needed would be over 500. Based on observing the sampling in Rochester Hills the previous day, we knew that it would not be possible to examine 500 ballots before the close of business day. To reduce the sample size needed, we reduced the sampling frame: I removed about 30 Election Day precincts from the ballot manifest and the reported vote totals.

(Another approach could have been to draw a fixed number of ballots, say 100, from the entire population and simply measure the attained risk of the sample. This approach would not have attained the desired 9% risk limit, but would have been equally illustrative.)

This changed the fraction of ballots with a CVR from 23.6% to 50%, reducing the workload estimate for hybrid SUITE. This did not substantially affect the reported margin between Ward and Neal amongst Election Day votes. The initial sample size estimate decreased to 260, with 130 ballots from the absentee votes and 130 from Election Day votes.

Lansing used 6 two-member audit boards. It took about 4 hours to pull and record all the ballots, then another hour to compare AV ballots to the CVR and tally the Election Day sample. The audit boards used counting down to sample AV ballots, because AV ballots were linked to their corresponding CVR by position: the batches of AV ballots and the CVR file were both supposed to contain ballots in the order in which they were scanned. They used a mix of counting and $k$-cut to sample Election Day ballots.

Among the 130 sampled AV ballots, we identified 15 discrepancies between the paper ballots and the CVR. It is unlikely that these were true discrepancies; we believe that the sampled paper ballots were matched to the incorrect CVR because of counting errors in sampling the ballots or the way the batches were stored. With so many AV discrepancies, the SUITE risk was 100%.

For comparison, we pooled the AV and Election Day samples for a pure ballot-polling audit. One AV ballot was sampled three times, reducing the overall ballot polling sample size to 258 ballots. In total, there were 116 votes for Ward, 94 votes for Neal, and 48 invalid ballots or write-ins. The initial sample size estimate for ballot polling was 285 ballots, just higher than the sample size actually drawn. Overall, the risk calculated using SUITE ballot polling was 87%. (The risk calculation using BRAVO ballot polling is 38%. This difference likely occurs because there were so many votes for neither Ward nor Neal in the sample.) In a true RLA, this would have led to a second round of sampling ballots.

**Kalamazoo**

The Kalamazoo pilot RLA took place on December 5. We audited the governor's race,

pretending that the governor's race was entirely contained in Kalamazoo city. In this contest, Gretchen Whitmer beat runner-up Bill Schuette by a margin of 54%.

The SUITE tool estimated that we would need to pull 33 ballots to achieve a 5% risk limit. Based on our experiences the previous two days, we knew this would not take long.

To ensure that we would meet the risk limit and to give each audit board extra hands-on experience, we increased the total sample size to 40 ballots, 8 from the AV ballots and 32 from Election Day. It took 4 two-member audit boards about 1.5 hours to both sample and tally the ballots.

The AV ballots were imprinted with ID numbers, so the comparison audit involved

looking up the imprinted ID in the CVR file. They did not rely on preserving the order of ballots as in Lansing, so a mix of counting and $k$-cut were used to sample both AV ballots and Election Day ballots.

There were no discrepancies in the AV sample. In the Election Day sample, there were 23 ballots for Whitmer, 8 ballots for Schuette, and 1 for Gelineau. The risk between Whitmer and Schuette was 3.7%, while the risk between Whitmer and the remaining

runner-ups was below 1%.

**Lessons Learned**

Ballot-level comparison audits are the most statistically efficient type of RLA: the number of ballots needed to end the audit, if the reported margin is correct, decreases inversely with the margin. (In comparison, the average sample size needed for ballot polling decreases inversely with the *square* of the margin.) It should be advantageous to use ballot-level comparisons when possible. However, we found empirically that when the majority of ballots do not have a CVR, then using ballot polling alone can be more efficient than using the hybrid SUITE RLA. In Lansing and Kalamazoo, the proportion of ballots with a CVR was below 25% and the hybrid RLA required sampling a few more ballots than ballot polling alone would have.

RLAs are meant to be flexible, and all that is absolutely necessary to conduct an RLA is a trustworthy paper trail. The choice of auditing method comes down to efficiency. Election officials should weigh efficiency of the statistical method, the number of ballots that a method requires to be examined, against the efficiency or feasibility of the process. For instance, we learned that for Lansing, ballot polling might be the most efficient option: ballot-level comparison is possible but difficult and SUITE is not necessarily more efficient than ballot polling.