Orange County Voter Registration Audit Report

R. Michael Alvarez and Silvia Kim
California Institute of Technology

October 25, 2018

1 Audit Summary: October 12, 2018

We have developed a methodology to audit the Orange County voter registration database for unusual activity. This change detection algorithm looks for record changes, records added, and records dropped. It produces a detailed report which we provide to the Orange County Registrar of Voters (OCRV) on a regular basis. Between April 26, 2018 and the most recent data we have received on October 12, 2018, our algorithm and interquartile range (IQR) analysis have found 28 'events.' These events all appear to be the result of normal database maintenance activities by OCRV.

2 Brief Description of Methodology

We have been examining daily changes in the Orange County voter database, starting April 26, 2018. Our basic methodology begins by noting that we have a previous snapshot of voter registration (VR) data (dataset $t-1$). We also have a current snapshot of VR data (dataset t).

Dataset $t-1$ will be split into three sub-datasets:

1. There are records that match exactly between previous data ($t-1$)

and current data (t), in all fields. We call these *exact matches* (usually 99% or more), as opposed to *unmatched* from $t-1$.

2. There are records that can only be found in $t-1$, and not t. These are *dropped records*.

3. There are records that are found in both data (identified using record linkage with partial string matching, or ID-only matching), but are changed between $t-1$ and t at some fields—for instance, address. ($t-1$ version). (5) and (6) make up the *"unmatched from $t-1$."*

Similarly, dataset t will be split into three sub-datasets:

4. There are records that are exactly the same between previous data ($t-1$) and current data (t).

5. There are records that can only be found in t, and not $t-1$. That is, these are *added records*.

6. There are records that are found in both data, but are changed between $t-1$ and t at some fields. (2) and (3) make up the *"unmatched from t."*

We distinguish between (2) and (3), and between (5) and (6), using record linkage techniques. This enables the linking of two records that are unmatched but point to the same entity. This happens because a person may re-register or update their information after moving, marrying, or similarly make changes to their information.

For the record linkage, we use the following selection of important variables (ADGN variables from Ansolabehere and Hersh, 2017) in the partial string matching.

- Name: szNameLast, szNameFirst
- Date of Birth: dtBirthDate
- Address: sSitusZip, sHouseNum

Matching is performed by `fastLink` (Imai, Enamorado, and Fifield, 2018). Default partial string matching parameters are employed. That is, 85% is the lower bound for the posterior probability of a match that will be accepted. Each column has the the lower bound for a partial agreement as 88%.

In addition to visual examination of the results of this process, we use a straightforward statistical anomaly detection method, the interquartile range (IQR) method. This aligns the data by size, and then finds points that fall outside of the $[Q1 - 1.5(Q3 - Q1), Q3 + 1.5(Q3 - Q1)]$ range, using the first quartile (Q1) and the third quartile (Q3).

3 Summary of the Results

Between April 26, 2018 and the most recent data we have received on October 12, 2018, our change detection algorithm and IQR analysis have found 28 events. These events all appear to be the result of normal database maintenance activities by OCRV, such as outlined in the 2018 Election Security Playbook, published by the OCRV. In the playbook's *Voter List Maintenance* section, the OCRV lists detailed plans for maintenance. We will update this report when we receive additional data.