Anomaly Detection and Automated Labeling for Voter Registration File Changes

Sam Royston
VoteShield, Protect Democracy
New York, USA
sam.royston@voteshield.us

Omeed Tavasoli
VoteShield, Protect Democracy
New York, USA
omeed.tavasoli@voteshield.us

Ben Greenberg
VoteShield, Protect Democracy
New York, USA
ben.greenberg@voteshield.us

Courtenay Cotton
VoteShield, Protect Democracy
New York, USA
courtenay.cotton@voteshield.us

1 INTRODUCTION

Statewide Voter Registration Files (VRFs) contain information about every voter within a state and determine who is allowed to participate in elections. As referenced in the U.S. Senate intelligence committee report released in Summer 2019, the infrastructure for managing these lists exposes a critical vulnerability in our electoral system [23]. By periodically taking snapshots of these files and analyzing the measured changes, we aim to automatically detect the nature of any modifications and flag those that require further review, adding important safeguards to the VRF maintenance process.

In this paper we present two main approaches to understanding the changes made to a voter file over time. In the first, we consider the (normalized) volume of different types of file changes as distributed over different locations and times. We apply several anomaly detection methods on these distributions in an unsupervised setting. We apply basic statistical methods to detect outliers with respect to both time and location. We also apply low-rank models, exploiting correlations in the data in order to suggest less obvious anomalies.

In our second approach, we broaden our model to consider the characteristics of those voters whose records were modified. Using a combination of demographic and historical features, we describe groups of changed voters using their aggregate features. We believe these individual voter features should be informative about the type of change that happened to those voters and the reason for that change, assuming it was a valid maintenance action. We examine this assumption with labeled examples of several known causes of a particular change type (deactivation), and use these features to predict the labels. The ability to automatically distinguish among maintenance actions with known causes is another step forward in distinguishing the set of expected voter file changes from unexpected, potentially malicious changes.

Anomaly detection methods are difficult to evaluate since the problem domains are often missing concrete examples of anomalies to check against, and our case is no exception. We only have access to a few anecdotal examples of erroneous deactivations and thankfully no examples of malicious modifications to the Voter File. The space of hypothetical attacks to the Voter File is not easily modeled, for instance because an attack would likely be selected from a set...
of diverse strategies, might occur only once, and because the objectives of an attacker could range from simply disrupting an election, to affecting the outcome. Further study, which is out of the scope of this work, is needed to model the many plausible types of attack; this is a main reason why in section 5 we focus on classifying the causes that we are familiar with, rather than the ones we are not. While various methods have been applied to detecting suspicious election results in the political science literature, the focus of our work is specifically on monitoring changes to voter registration databases leading up to an election. To our knowledge, the similar previous work has either not been applied in production, or has not focused specifically on using anomaly detection or supervised learning to help protect and maintain voter databases. Our contribution is a description of which anomaly detection methods work in this applied setting, providing baselines for future research, as well as a proposed automatic labeling system to help further reduce the burden on administrators managing these databases.

2 PRIOR WORK

In the field of political science, there has been some academic study of voter registration databases, but limited previous research into the dynamics of changes to these databases. Older works such as [2], [5], and [3] establish some baseline analyses of the data quality or lack thereof (in terms of availability, consistency, completeness, and accuracy) of voter registration data and voting records across states in the U.S. In [18] Merivaki performs a case study in Missouri, modeling voter data error rates and suggesting a relationship between error rates and a county’s voter status data and other demographic data.

The Electronic Registration Information Center (ERIC) project [6] is a non-partisan, non-profit organization which facilitates data sharing and analysis for its 29 member states. ERIC software matches duplicate voters on different states’ VRFs so states can more effectively rid their voter files of stale records, in addition to a number of other useful administrative functions. We find ERIC to be the most similar project to ours which is deployed and in active use, though their approach and functionality is based almost entirely on record linkage, rather than detection of anomalous changes.

More recently, there have been some attempts to model how voter registration databases change over time and identify when changes are anomalous. In [21] Pettigrew and Stewart attempt to model and evaluate the accuracy of voter removal processes only, due to death or moving. In a case study of two states, they highlight the fact that different states have very different voter list maintenance procedures.

Later in [15] Kim, Schneider, and Alvarez present a framework somewhat similar to the one we will present in section 4.2. They take many snapshots of a voter database over a period of time and identify the changes between them, in terms of records removed, added or changed. They suggest using an inter-quartile range (IQR) threshold applied to raw change counts across many snapshots in a single county in Florida to identify anomalies over time. Furthermore, they confirm with the county’s registrar that the discovered anomalies in fact correspond to specific infrequent maintenance events. This contrasts with our work since one of our objectives is to identify anomalies that cannot be classified as routine maintenance, regardless of how infrequently some maintenances may occur. In [7] two of the same authors extend this work to modeling all counties in Florida over time. They derive parameters for a Bayesian model relating snapshot time, location (county), and county population size to voter record change rates. The authors then define anomalies as points with significant deviation from this model’s predictions.

Anomaly detection models have also been applied in political science on election outcome data itself [16, 28]. More broadly, anomaly detection is a well-studied problem in many domains [8]. The problem we study here bears some resemblance to fraud detection in financial domains [1] and database security problems [14]. The techniques we use here range from straightforward statistical methods to matrix decomposition methods, which have often been applied to anomaly problems [19, 25, 27]. In section 5 we also present a classification task that we intend to use ultimately to aid in identifying anomalies, which also has precedent [22, 24], though not in the realm of VRF maintenance.

3 VOTER REGISTRATION FILES

Voter Registration Files contain information about every voter who is registered in the state and sometimes voters who were registered in the past but have since been unregistered. These files contain different information depending on the state, but generally all contain a set of fields describing each voter’s name, address, age, political party, voter status, and history of participation in past elections. We restrict our analyses to the voter file from the State of Iowa, though files from other states contain roughly the same information and there is no reason to believe that a variant of these methods would not be applicable in other states. Only a subset of the fields describing each voter affect their ability to vote, including: address fields, name fields, activation status, party affiliation and of course their presence on the list in the first place (their removal or registration status). We extract changes to these fields by computing differences between temporally adjacent versions of the file. The change types that we use in our analyses are defined in table 1. The modification data are inserted into a database and later aggregated by change date, locale (county), and change type for further analysis and anomaly detection, covered in sections 4 and 5.

4 ANOMALY DETECTION

This section deals with analyzing change counts when measured in aggregate by locale. 1 Treating each (date, locale) pair as a single aggregate sample instead of dealing with individual records has a few benefits:

Anonymity: Input data do not contain any personally identifiable information.
Dataset size: Aggregating by locale reduces computational load.
Actionable results: There will be a responsible party for addressing any issues found.

1 A locale is a geographic administrative unit, like a county
Table 1: Types of Voter Registration File changes which can affect someone’s ability to vote

| Change Type | Description |
|-------------|-------------|
| Address     | Changes to address fields, for example house-num, street-name, etc. |
| Name        | Changes to first-name, middle-name, or last-name fields. |
| Removal     | A record is deleted (in the comparison no matching voter-id is found in the posterior file) |
| Registration| An insertion of a new record (no matching voter-id is found in the anterior file) |
| Deactivation| A change in the voter-status field from active to inactive. Inactive voters can still vote, but are eligible for removal in maintenance processes. |
| Activation  | A change in the voter-status field from inactive to active |
| Party       | A change in the party-affiliation field |

Moreover, administrative units are also the lens through which an attacker would view access to the VRF. Different counties often use different database technologies than one another and have different security practices. In the case of a hack, a likely target could be groupings of administrative units which have certain infrastructure configurations in common. These sub-populations are assembled by aggregating modifications to the VRF by the date, administrative unit, and change type. For example, all address changes in Polk County measured between January 3rd and January 10th, 2019 would constitute a sub-population of changes.

In the later sections 6.2 and 6.3, we report these methods’ performance in detecting sparse artificial changes, as well as real-life examples of how they performed when applied to deactivations in the state of Iowa over a 2.5-year period spanning 2018 to 2020. In this section we describe the rationale of each of these methods in detail.

4.1 The Modification Matrix

We will make frequent use of the modification matrix, denoted $M$; a matrix of scalar values associated with each sub-population. All techniques discussed in this section operate on this matrix, whose entries are a normalized count of changes associated with each date and locale, for a given change type. The normalization used is changes per-day, per 1000 voters. Consider the modification matrix $M \in \mathbb{R}^{\ell \times d}$ where $\ell$ is the number of locales and $d$ is the number of time intervals (rendered in figure 1). Each column is a different time window over which modifications were tabulated and each row is a distinct locale. We refer to administrative boundaries as locales, but any partitioning of the electorate can be used to populate the rows of $M$, and we suspect that alternative partitions may also be useful.

4.2 Temporal Method

This mode of anomaly detection stages the problem as one-dimensional time-series analysis. Since we do not have sufficient data to span multiple election cycles we must eschew seasonal models, and have no reason to believe that an auto-correlative assumption is valid here. Based on this logic, we choose to evaluate simple univariate descriptive statistics based on the Inter Quartile Range and standard deviation.

$$IQR(x) = \frac{Q_3 - Q_1}{\sigma}$$

$$zscore(x) = \frac{x - \mu}{\sigma}$$

where $\sigma$ is the standard deviation, $\mu$ is the arithmetic mean, and $Q_n$, $n \leq 4$ is the $n$th quartile. In temporal anomaly detection, we compute our descriptive statistics across each successive date range for each locale, e.g. there is a unique set of $Q$s along with $\mu$ and $\sigma$ for each locale (and change type). In our illustrations the row order is alphabetical, but this ordering is never relevant for the techniques we propose. For temporal methods, $(\sigma, \mu, Q_n)$ are computed individually for each row, and does not change the order from if we just ranked the raw per-day, per-1000 voters score. It’s worth noting that the IQR method from [15] falls into this category.

In practice, the temporal method picks out county, date pairs primarily during the routine mailings of address confirmation postcards that took place in February of 2019 (C). This is a useful baseline which mimics what a human might do when inspecting a time series chart (one would pick out the spikes). One would not want
to rely exclusively on this method since it is predisposed to identify only large scale changes that are part of routine maintenance.

4.3 Cross-Locale Method

When normalized by population, the maintenance behavior should be roughly identical (i.e. appear to draw from the same distribution) across all locales at a single point in time. Given this assumption, it makes sense to detect outliers by comparing behavior across different locales within a restricted time range. We use the same metrics listed in (1) but instead of computing \((\sigma, \mu \text{ or } Q_n)\) per row, we compute them over a sliding set of columns of width \(2w + 1\) centered on the time interval corresponding to \(x\). The purpose of selecting \(w > 0\) is to account for processes which may not co-occur exactly, but do co-occur within a limited amount of time.

For \(M_t \subset M, M_t \in \mathbb{R}^{\ell \times (2w+1)}\) and with respect to \(x\) we have,

\[
M_t(x) = \begin{pmatrix}
\begin{array}{cccc}
 a_{1,col(x)-w} & \cdots & a_{1,col(x)+w} \\
 \vdots & \ddots & \vdots \\
 a_{\ell,col(x)-w} & \cdots & a_{\ell,col(x)+w}
\end{array}
\end{pmatrix}
\]

Combining this definition with the IQR or Z-Score from (1), yields a complete description of the Cross-Locale method. The IQRs and Z-Score\(^2\) methods yield similar results; the IQR method is superior when multiple anomalies co-occur temporally, while the Z-Score method is more sensitive to anomalies of lower magnitude. In our implementations we found that \(w = 2\) worked the best, which corresponds to sliding windows with a width of about 5 weeks. See CL_IQR_5, CL_IQR_3, CL_STD_5, and CL_STD_3, in figure 8 and note that CL_IQR_N corresponds to the full window width \(N\) where \(N = 2w + 1\).

Qualitatively, this detection method is less activated when spikes happen en masse across county (e.g. events B and C) and instead will highlight events that seem out of place compared to what else is going on at that point in time. Figure 2 shows an example of

\(^2\)the terms Z-Score and STD are used interchangeably.

4.4 Low-Rank Models

Modeling normal changes as a low rank matrix offers a more sophisticated approach to anomaly detection on \(M\) than assuming each row or column represents an independent distribution of samples like we do in sections 4.2 and 4.3.

The top singular values seen in figure 3 show us that these modification matrices are approximately low-rank (i.e. they have many singular values close to zero), so it makes sense to use models that rely on a low-rank assumption to model the different types of systematic maintenance.

4.4.1 Non-Negative Matrix Factorization. Non-negative matrix factorization \([9, 10, 12]\) involves finding a factorization of \(M \in \mathbb{R}^{\ell \times d}\) consisting of two factor matrices \(P \in \mathbb{R}^{\ell \times k}\) and \(Q \in \mathbb{R}^{d \times k}\). The NMF anomaly detector assumes that the legitimate components of \(M\) are the product of a limited number of factors associated with each date and locale and allows explicit control over the number \(k\) of latent factors. Put another way, we can use \(k\) to restrict the presumed rank of the matrix resulting from a legitimate process.
This method also takes advantage of the non-negativity of \( M \); since \( M \) contains only normalized counts, all entries must be zero or greater. This allows for restricting the factor matrices to be non-negative as well, resulting in fewer possible solutions to the optimization and (usually) better results [9]. \( P, Q \) are selected to minimize the squared Frobenius norm (squared element-wise \( \ell_2 \) matrix norm) of the difference between \( M \) and the factored reconstruction \( PQ^T \).

We select \( P, Q \) so as to minimize

\[
\minimize \| M - PQ^T \|_2^2 \\
\text{subject to } P > 0, \ Q > 0
\]

(2)

and take the largest entries from \( M - PQ^T \) to be the most anomalous. Figure 4 shows as example of this difference output. We use the implementation in [20] which uses the coordinate descent method described in [12] to solve (2).

4.4.2 Robust Principal Component Analysis. Robust Principal Component Analysis [11] attempts to decompose an input matrix into the sum of two matrices, one low-rank and one sparse. This formulation is uniquely appropriate since it seeks to model the underlying legitimate process and the anomaly process separately. We perform anomaly detection by inspecting the highest-value entries of the sparse matrix. Utilizing RPCA relies on applying a few generalizations about the structure of our modification matrix \( M \).

- The matrix of natural changes to the VRF is approximately low-rank.
- There are a small number of outliers within \( M \) which do not adhere to this low-rank structure, i.e. the matrix of outliers is sparse.
- The original matrix of changes is the sum of one natural and one outlier matrix

\[
M = L + S
\]

Where \( S \) is sparse and \( L \) is low-rank. Finding such matrices is equivalent to solving the following optimization

\[
\minimize ||L||_* + \lambda ||S||_1 \\
\text{subject to } M = L + S
\]

(3)

Where \( ||L||_* \) is the nuclear norm of \( L \) [26]; a proxy for minimizing the rank directly (the latter is known to be computationally intractable), and \( ||S||_1 \) is the entry-wise \( \ell_1 \) norm of \( S \). We used the implementation of the Iterative Shrinkage Thresholding Algorithm [4] found in [13] to solve equation 3 and perform anomaly detection by ranking the entries of \( S \).

5 AUTOMATIC LABELING USING POPULATION METADATA

The techniques discussed in section 4 rely on structural properties of the modification matrices to highlight certain abnormal (date, locale, change type) groups in terms of a single metric: changes per day, per one thousand voters. These techniques make no distinction between modifications which affect different demographic groups within a locale’s population.

In this section, we present a framework for predicting anomalies using population metadata. Incorporating features such as age, electoral participation, and history of address changes has allowed us to observe statistical regularities at the population level, and identify as potential anomalies the (date, locale, change type) groups which differ from the model’s expectation. The full set of features we use is listed in table 2.

With access to VRF snapshots, we can use demographics, voting history, and change history data in order to predict anomalies. We aggregate voter data at the level of change groups; as in the modification matrix, our basic unit of population is the set of voters with modifications of some type, within some locale, between two consecutive snapshots. We compute the means of demographics- and history-derived features for the voters in each change group. These features may allow us to draw conclusions about the cause of the changes.

5.1 Population Metadata

The availability of demographic variables varies by state – the Iowa voter file tracks birthday, gender, voter status, and party identification. We treat age (in years) as a continuous numerical variable, and the others as one-hot encoded categorical variables. (E.g. for voter status, we use three features: voter status active, voter status inactive, and voter status pending, with binary values representing the status of each voter).

We use features derived from individual voting histories, including numerical measures of turnout: participation (how often a voter turned out for elections in which they were eligible), partisanship (how often a voter turned out for their party), engagement (measured as the average size of elections a voter participated in), and provisional vote count. We also track temporal features, including days since last voted, months since registration, and months since the last VRF update.

Finally, the change history features measure how voters in a change group have been affected by each type of modification in the past. For each of the seven change types (see table 1), we counted the number of changes affecting each voter during the prior 6 months (up to and not including the “current” change), as well as “since the first VRF snapshot” (again, not including the current change).

Concretely, for \( n \) demographic features associated with each voter \( v \in \mathbb{R}^n \) we compute \( \psi_p(v) \in \mathbb{R}^n \) which is the column-wise mean of each voter within a single sub-population \( p \in P \) where \( P \) is the set of population partitions, defined by the date, locale, and change type. Figure 5 shows a t-Stochastic Neighbor Embedding created with this feature representation of (date, locale, change type) populations with a few interesting clusters highlighted, including the general region of some of the labels used by our supervised model (table 4). The clear divisions between change types help validate the usefulness of these features.

5.2 Predicting Event Labels

Many modifications occur due to so-called “systematic processes” – semi-regular VRF changes affecting a subset of voters in a locale. These processes define a pattern of normal behavior for the databases, and the corresponding modifications should not be flagged as anomalies. Therefore, we would like to predict the systematic process affecting a group of voters, using demographic and history data, in order to avoid false positives in anomaly detection.
Figure 5: Plot of a t-Stochastic Neighbor Embedding (t-SNE) [17] of the features described in section 5.1, colored by their change type. Labels are: (A) NCOA or Similar Mailing - Response Processing: 2019/2020 annual deactivations of non-respondents to election-related postcards (B) Systematic September Maintenance (C) Forwarded NCOA Mailing - No Response (D) Inactivity Removals: Removal of voters due to inactivity (as opposed to death or incarceration) (E) Libertarian Party Reclassification: Party changes caused by the reclassification of the Libertarian party to a political organization

Table 2: Demographic and history features for Iowa voters

| Variable                           | Type     | Class          |
|------------------------------------|----------|----------------|
| months since registration          | numerical| demographic    |
| years old                          | numerical| demographic    |
| gender                             | categorical| demographic   |
| voter status                        | categorical| demographic   |
| party membership                   | categorical| demographic   |
| days since last voted              | numerical| voting history |
| partisanship                       | numerical| voting history |
| participation                      | numerical| voting history |
| engagement                         | numerical| voting history |
| provisional votes                  | numerical| voting history |
| absentee votes                     | numerical| voting history |
| <change type> changes last 6 months| numerical| change history |
| <change type> changes all time     | numerical| change history |

Systematic processes are periodic VRF updates, initiated by state or local election officials in order to maintain an accurate voter file. For example, a state might periodically remove convicted felons from the voting roll as arrest records become available, or deactivate voters who fail to respond to a National Change of Address (NCOA) mailing.

For each change group, we calculated mean values of each demographic, voting history, and change history feature. After transforming to have zero-mean and unit-variance for the input, we trained an XGBoost classifier to predict which of the event labels would most likely explain the changes of the (locale, date) group in question.

A robust label classifier can be used to complement an existing anomaly detection scheme. If the classifier predicts with high confidence that a set of modifications occurred due to a known pattern of behavior, an anomaly detection system should assign a lower probability that the modifications are anomalous. This creates a positive feedback loop; new user-provided event labels are used to train the classifier, which is in turn used to recommend examples in need of labeling.

6 EXPERIMENTAL EVALUATION

6.1 Improper Changes Modeled as Sparse Additive Noise

To our knowledge there exist few to no comparative evaluations of multiple anomaly detection techniques for VRDBs across the different types of changes which occur to VRFs, which we provide here.

In this section we evaluate our methods for detecting unexpected modifications in a modification matrix (described in section 4). To do this, we perturbed 1% of the entries at random in the modification matrix by adding a positive component \( \gamma \). For each change type, we tested each method’s ability to discern which entries were modified by treating them as a ranking system: we took the top 20 highest ranking entries and reported the proportion of those which were members of the perturbed set. As \( \gamma \) becomes larger, the task becomes easier and easier, and for each change type we tested the methods over \( \gamma \) ranging from 0 to 20 times the average daily changes per 1000 voters. We based the \( \gamma \) ranges on the averages to mitigate the effects of change frequency on the reported scores; otherwise for example, name changes would have appeared to be significantly easier to monitor than address changes because they are rarer. Methods which achieve a higher precision at lower \( \gamma \) are better at detecting this type of modification, and we use AUC to summarize each method’s performance over the range of \( \gamma \).

It is important to recognize the major simplifications present in this evaluation. For one, we knew \textit{a priori} that there were greater than 20 perturbed entries, therefore the choice of selecting the top 20 suggested by the candidate methods is an artificially safe one. In a real world use case we would not know much about the size of the perturbed set. For this reason, these evaluations should be seen as a coarse affirmation of the ability to recommend relevant anomalies for review, rather than an affirmation of our ability to detect anomalies.

Also included as baselines are global versions of the \( STD \) and \( IQR \) methods: GLOBAL STD and GLOBAL IQR, respectively. These methods calculate the group statistics over the entire \( M \), and since they both apply a single linear function to all values in \( M \), they will both result in the same rankings as the raw input values. In addition to these baselines, we included our methods based on NMF
Similar to the inactivity mailing above, with deactivations window width of 3

**Table 3: AUC for each change type and strategy, for γ ranging from 0 to 20x the average daily changes per 1000 voters for the relevant change type.** This table is intended to facilitate comparisons between modification-ranking methodologies.

| Event Label     | Description                                                                                     | Count |
|-----------------|-------------------------------------------------------------------------------------------------|-------|
| Inactivity Mailing - Response Processing | A bulk deactivation due to processed responses (or lack thereof) from mailings sent by the government to ensure the veracity of address occupancy records | 99    |
| Systematic September Maintenance | An unspecified deactivation maintenance procedure which took place around Sep. 2019 | 37    |
| NCOA Mailings | Similar to the inactivity mailing above, except that the mailings were sent as part of the maintenance of the National Change Of Address database | 27    |
| Other | Any deactivation not included in the above classes, often composed of only a small number of changes | 21    |

**Table 4: Event labels for systematic processes associated with deactivations**

| Event Label          | Description                                                                                     | Count |
|----------------------|-------------------------------------------------------------------------------------------------|-------|
| Deactivations        | **NMF 0.38**                                        | 0.23  |
|                      | **RPCA 0.59**                                      | 0.41  |
|                      | **CL IQR 3 0.72**                                 | 0.66  |
|                      | **CL STD 5 0.66**                                 | 0.03  |
|                      | **CL STD 3 0.03**                                 | 0.02  |
|                      | **TEMPORAL IQR 0.03**                             | 0.04  |
|                      | **TEMPORAL STD 0.04**                             | 0.02  |
|                      | **GLOBAL IQR 0.02**                               | 0.02  |
|                      | **GLOBAL STD 0.02**                               | 0.02  |
| Address              | **NMF 0.59**                                        | 0.41  |
|                      | **RPCA 0.59**                                      | 0.41  |
|                      | **CL IQR 3 0.76**                                 | 0.33  |
|                      | **CL STD 5 0.33**                                 | 0.38  |
|                      | **CL STD 3 0.38**                                 | 0.18  |
|                      | **TEMPORAL IQR 0.18**                             | 0.17  |
|                      | **TEMPORAL STD 0.18**                             | 0.19  |
|                      | **GLOBAL IQR 0.19**                               | 0.27  |
|                      | **GLOBAL STD 0.28**                               | 0.27  |
| Name                 | **NMF 0.66**                                        | 0.41  |
|                      | **RPCA 0.59**                                      | 0.41  |
|                      | **CL IQR 3 0.63**                                 | 0.66  |
|                      | **CL STD 5 0.66**                                 | 0.3  |
|                      | **CL STD 3 0.31**                                 | 0.31  |
|                      | **TEMPORAL IQR 0.31**                             | 0.32  |
|                      | **TEMPORAL STD 0.31**                             | 0.33  |
|                      | **GLOBAL IQR 0.33**                               | 0.33  |
|                      | **GLOBAL STD 0.28**                               | 0.28  |
| Removal              | **NMF 0.53**                                        | 0.26  |
|                      | **RPCA 0.59**                                      | 0.4  |
|                      | **CL IQR 3 0.52**                                 | 0.51  |
|                      | **CL STD 5 0.51**                                 | 0.14  |
|                      | **CL STD 3 0.14**                                 | 0.35  |
|                      | **TEMPORAL IQR 0.06**                             | 0.09  |
|                      | **TEMPORAL STD 0.09**                             | 0.08  |
|                      | **GLOBAL IQR 0.08**                               | 0.08  |
|                      | **GLOBAL STD 0.32**                               | 0.32  |
| Activations          | **NMF 0.74**                                        | 0.43  |
|                      | **RPCA 0.59**                                      | 0.4  |
|                      | **CL IQR 3 0.83**                                 | 0.63  |
|                      | **CL STD 5 0.63**                                 | 0.6  |
|                      | **CL STD 3 0.37**                                 | 0.37  |
|                      | **TEMPORAL IQR 0.43**                             | 0.38  |
|                      | **TEMPORAL STD 0.43**                             | 0.39  |
|                      | **GLOBAL IQR 0.39**                               | 0.39  |
|                      | **GLOBAL STD 0.33**                               | 0.33  |
| Registrations        | **NMF 0.74**                                        | 0.05  |
|                      | **RPCA 0.66**                                      | 0.39  |
|                      | **CL IQR 3 0.84**                                 | 0.84  |
|                      | **CL STD 5 0.84**                                 | 0.06  |
|                      | **CL STD 3 0.36**                                 | 0.38  |
|                      | **TEMPORAL IQR 0.06**                             | 0.06  |
|                      | **TEMPORAL STD 0.06**                             | 0.05  |
|                      | **GLOBAL IQR 0.07**                               | 0.07  |
|                      | **GLOBAL STD 0.38**                               | 0.38  |
| Party                | **NMF 0.66**                                        | 1.0   |
|                      | **RPCA 0.66**                                      | 1.0   |
|                      | **CL IQR 3 0.83**                                 | 1.0   |
|                      | **CL STD 5 0.84**                                 | 1.0   |
|                      | **CL STD 3 0.36**                                 | 1.0   |
|                      | **TEMPORAL IQR 0.06**                             | 1.0   |
|                      | **TEMPORAL STD 0.06**                             | 1.0   |
|                      | **GLOBAL IQR 0.07**                               | 1.0   |
|                      | **GLOBAL STD 0.38**                               | 1.0   |

**6.2 Event Label Classification Results**

We trained a classifier to predict event labels for the voter features described in section 5.1 computed for deactivations in Iowa from June 2018 to Dec 2020. Deactivations are a step in the process of removing voters from VRFs who have not voted recently or did not respond to validation mail sent by election administrators. We used this change type for our labeling experiment since it usually only happens in large groups for specific reasons (in contrast to name changes for example, which usually occur organically at the request of the voter). We hand-labeled 184 examples by examining the conditions of each deactivation group and validating our assessment with the office of the Iowa Secretary of State. Table 4 describes the four labels we applied and their counts in the dataset.

For each of the 184 groups of modified voters, we computed the demographics, voting history, and change history features using historical data. 20% of the groups were reserved for the holdout set; the rest were used as inputs for training an XGBoost model with 50 estimators of depth 3, with a learning rate of 0.3.

Of the four labels, one (NCOA or Similar Mailing - Response Processing) accounted for 99 of the 184 labels. We achieved a classification accuracy of 0.892, with an F1 score of 0.892, as seen in Table 5. Table 6 shows the full confusion matrix from this experiment. At this level of accuracy, we are able to correctly identify the dominant event type affecting most groups of deactivations. As we only collected 184 labeled groups for this experiment, we hope that a larger dataset covering more locales, dates, and event types will make this classifier more effective for anomaly detection. Furthermore, we expect that predicting labels from modifications at the level of individual records, as opposed to population groups, will yield a more finely-grained event label classifier.

**Table 5: Classification accuracy and F1 score on the holdout set for the label classifier**

| Accuracy | F1-Score |
|----------|----------|
| 0.892    | 0.892    |

**6.3 Identifying Real-Life Improper Deactivations**

Due to a misunderstanding related to a school board election mailing, four counties in Iowa inadvertently deactivated several hundred county records
When features from these four change groups were fed into the event label classifier described in section 5.2, they were predicted to have the label \textit{NCOA or Similar Mailing - Response Processing} with probabilities 0.995, 0.990, 0.987, and 0.952. These would be quickly labeled as anomalous, as NCOA mailings are generally processed during large, statewide maintenance events.

7 CONCLUSION

In this paper we present two broad models for understanding the dynamics of changes made to voter registration files. Both models begin with the same high level characterization of voter file changes as groups defined by their date, locale, and change type.

Our first type of model describes these change groups with a single normalized count feature. For each change type, these counts compose a modification matrix (of locales and dates) that can be operated on statistically or by using matrix decomposition methods to detect different types of unexpected values. We validated these models by adding random noise to certain matrix entries and evaluating the precision of each model in detecting those modified entries. Our NMF method and our Cross-Locale methods performed best on this synthetic noise task. Additionally, we have evidence from a small number of known real-life anomalous deactivation groups that our NMF, Cross-Locale Z-Scores, and RPCA methods would perform well in detecting those anomalies (and in fact did detect them in real life).

Our second class of model describes the same change groups with a more rich and descriptive feature set, taking demographic and history feature averages of the changed population. We evaluated this model on the more realistic and supervised task of classifying among four known reasons for voter deactivations, and were able to correctly predict these reasons with 89.2% accuracy.

Both model types may be applied to the core problem of identifying anomalous changes in a VRF. The unsupervised models can be applied directly to rank change groups as potential anomalies, while the supervised model can help us learn common change patterns and thereby direct attention toward changes that do not fit any known pattern. In the future, we plan to apply a similar labeling model to record level data such that we can predict the cause for each individual change to the VRF as opposed to causes associated with aggregate changes. We hope that our work inspires further interest in using machine learning to aid in the administration of Voter Registration Files.

ACKNOWLEDGMENTS

We would like to thank Carlos Fernandez-Granda of NYU for his valuable advice and feedback. We would also like to thank Wes Hicok and the Iowa Secretary of State’s Office, and Tim Hagle of the University of Iowa, for their help in understanding and labeling this data and for their feedback as early users of our anomaly detection system. Lastly, we would like to thank the entire VoteShield team for making this research possible.

REFERENCES

[1] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. 2016. A survey of anomaly detection techniques in financial domain. \textit{Future Generation Computer Systems} \textbf{55} (2016), 278–288.
Figure 8: Precision in identifying artificial changes applied to 1% of entries of the modification matrix, plotted as a function of $\gamma$, the magnitude of the artificial changes for each change type. The domain of each curve is always $[0, 20\mu M]$ where $\mu M$ is the mean number of changes per day per 1000 voters for the relevant change type. Therefore, since name changes occur less frequently than address changes, the corresponding curves appear truncated.

Table 6: Confusion matrix for event label classification. We saved 37 out of 184 labeled groups for the holdout set. Rows correspond to ground truth data; columns correspond to predicted labels.

|        | (1) | (2) | (3) | (4) |
|--------|-----|-----|-----|-----|
| (1) Response Processing | 18  | 0   | 1   | 0   |
| (2) September       | 1   | 8   | 0   | 0   |
| (3) No Response     | 0   | 0   | 6   | 1   |
| (4) Other            | 1   | 0   | 0   | 1   |