Heterogeneity in Voter List Maintenance Practices: A Study of Florida Counties

Jian Cao*, Seo-young Silvia Kim, and R. Michael Alvarez

California Institute of Technology

April 2, 2020

Abstract

How do we ensure the accuracy and integrity of a statewide voter registration database, which often depends on aggregating decentralized, sub-state data with different list maintenance practices? We present Bayesian multivariate multilevel model to account for common patterns in local data while detecting anomalous patterns, using Florida as our example. We use monthly snapshots of state's voter database to estimate countywide change rates for multiple response variables (e.g., changes in voter's partisan affiliation), and then jointly model their changes. We show that there is much heterogeneity in how counties manage voter lists, resulting in very different patterns in additions, deletions, or changes of records. Our method identifies several Florida counties with anomalous rates of changes in the 2016 election.

1 Introduction

In 2002, the *Help America Vote Act* (HAVA) was enacted in the United States, a sweeping election reform package that sought to resolve many of the issues that had plagued the 2000 U.S. presidential election. One of the important provisions in HAVA was Section 303, which required that each state "implement, in a uniform and nondiscriminatory manner, a single, uniform, official, centralized, interactive computerized statewide voter registration list defined, maintained, and administered at the State level that contains the name and registration information of every legally registered voter in the State and assigns a unique identifier to each legally registered voter in the State ..."[1]

*Cao is a postdoctoral researcher at Caltech, and is the corresponding author (email: jccit@caltech.edu); Kim is a Ph.D. candidate in the Division of Humanities and Social Science, Caltech; Alvarez is Professor of Political and Computational Social Science, Division of Humanities and Social Sciences, Caltech. The authors thank the John Randolph Haynes and Dora Haynes Foundation for supporting their research. SK and RMA conceived research; JC, SK and RMA designed research; SK and RMA acquired data; JC and SK processed and analyzed data; JC, SK and RMA interpreted data; JC, SK, and RMA wrote the paper.

[1]The *Help American Vote Act* is P.L. 107-252 (2002). Complete text is available at https://www.eac.gov/assets/1/6/HAVA41.PDF.

Within a few years after passage of HAVA, most states had implemented statewide voter registration databases, though their design and structure varied considerably by state (Liebschutz and Palazzolo, 2005).

Fast-forward nearly two decades, and concerns are increasing about the reliability, accuracy, integrity, and security of these legacy voter registration systems used in many states. Maintenance of voter data is extremely complex and demanding work, and sometimes things go awry even with the best intentions. In California, it was reported that the automatic voter registration caused a flood of duplicate records and registrations with the wrong party in 2019, with counties showing differential rates of voters registering as 'no party preference.' In Ohio, amid controversies of voter purge, differential purge rates were reported within the state, and Franklin County in particular was under fire for wrongly cancelling voter registrations. List maintenance issues like these might be considered as internal threats to the integrity of a voter registration database.

Furthermore, in both 2016 and 2018, media reports alleged that hackers may have tried to access state voter registration and election administration systems across the United States, reports that were confirmed by the "Mueller Report," which presented evidence indicating that hackers may have gained access to the election administration and voter registration systems in Illinois and in some Florida counties in 2016 (Mueller, 2019). These are external threats to the integrity of voter registration data. It is thus imperative that we develop tools to ensure that voter registration list maintenance practices produce accurate and reliable voter data for election administration (protecting against internal threats), as well as to ensure the security and integrity of the data (protecting against external threats).

It should be noted that external and internal threats to the integrity of data can manifest very differently by jurisdiction. This is because election administration is highly decentralized in the United States, and local variations exist. Therefore, the question we consider in this paper is the following: how do we ensure the accuracy and integrity of a statewide voter registration database, which depends on the aggregation of decentralized, sub-state databases? To put it another way, how can we detect potential anomalies in the voter registration data that could indicate errors, or which even might provide evidence of potential intrusion into the database systems—while taking into account the potential variance in how each jurisdiction manages their voter data? Because of this decentralization, and because there are often concerns raised about the integrity of voter registration data, this is a very important question that needs to be systematically addressed. Strong data integrity methods can detect anomalies if there are any, and help to mitigate concerns about data security and reliability, if there are no unintended "anomalies."

In this paper, using voter registration data from the state of Florida in the 2016 presidential election cycle, we extend the approach taken by Kim, Schneider and Alvarez (2019). This paper demonstrated the utility of analyzing daily snapshots of at single county's voter registration database for anomaly detection. While Kim, Schneider and Alvarez (2019) considered a single county and separately analyzed each of the variables in question—such as changes in addresses or changes in

party affiliation—here we extend this approach to pool the information across variables, and also different counties and change periods. This reflects the assumption that the latent factor producing change in one variable will also be reflected in changes of another variable, and thus the change rates may correlated between them. In addition, the changes will be affected by county-specific effects and change time period-specific effects. To express these, we perform a Bayesian analysis on monthly snapshots of an entire state's voter registration database, and use a model in which there are multiple response variables with group-level effects from counties and from time. The model also takes into account heterogeneity across counties such as population size. All of this allows a more principled modeling of why the database is changing, and thus a better assessment of what constitutes an anomaly.

The next section discusses the limited past research in this area, and gives an overview of our methodology. We then discuss the data we use and our model, followed by a presentation our results. We conclude with a discussion of the utility of this approach for analyzing anomalies in county and state voter registration databases.

2 Voter Registration Database Integrity

While concerns about the integrity of voter registration databases have been raised in the media and in public discussion, there has been scant academic research on how to measure the quality of voter registration data, and how to detect errors or evidence of possible intrusion into these administrative databases. Statewide and computerized voter registration lists were required by the *Help America Vote Act* (HAVA), which was passed into law in 2002. The rationale for HAVA was that the development and use of statewide computerized registration databases would produce more accurate and usable registration data, that might minimize many of the problems that had been observed with errors in voter registration in the 2000 presidential election (Caltech/MIT Voting Technology Project, 2001). While some academics raised concerns about the integrity and security of these large administrative databases (Caltech/MIT Voting Technology Project, 2001; Alvarez, 2005), it took nearly a decade for researchers to being to develop methodologies for confirming the integrity and accuracy of voter registration data.

There were various methods that became the subject of academic attention for validating voter datasets. An early effort linked voter registration data between Oregon and Washington, which are adjacent states, in an effort to look for potential duplicate records in both states (Alvarez et al., 2009). Others used third-party data—in particular, surveys of registered voters who were asked to confirm the information about them in the registration database—to examine the accuracy of information in voter databases (Ansolabehere and Hersh, 2010, 2014). Researchers using administrative data such as voter registration information, for other research purposes like studying turnout, also began to question the accuracy of the information in these large databases (Berent, Krosnick and Lupia, 2016; Green and Gerber, 2006).

More recently, researchers have begun to dig even further into the quality of voter registration databases. Some have looked at the problem from the perspective of voter list maintenance practices. Ansolabehere and Hersh (2010) has evaluated static data quality such as missing birth dates or addresses. Pettigrew and Stewart III (2017) investigated two different paradigms in removing voters who have moved out of the jurisdiction. Detection of duplicate records in voter registration datasets has also been the subject of some attention (Christen, 2012, 2014; Christen and Gayler, 2013).

More importantly, there has been more attention in the literature on how variations in list maintenance practices at the local level could create heterogeneity in statewide voter registration data quality, in particular in states with bottom-up or hybrid list maintenance practices. Merivaki (2019) investigates how rejections of voter registration applications vary locally, depending on the time of the year as well as registration sources. Also, Merivaki (2020) investigates how registration and voting history errors are dependent on local socio-demographics and inactive voter rates. These studies provide the foundation for the analysis we report in this paper.

Our approach is similar to Kim, Schneider and Alvarez (2019). We follow their approach of using repeated instances (or "snapshots") of a jurisdiction's voter registration database, and like their approach we use repeated record linkage for entity resolution between those snapshots. We then use the matched results to build multiple time-series that will help us assess the database quality, such as the time-series of number of new records, the number of records dropped, and number of records that changed in key fields (e.g. address or partisan affiliation) between the snapshots. Like Kim, Schneider and Alvarez (2019), we then use statistical anomaly detection tools to ascertain whether a particular rate of change in any of these time series are statistical outliers—and thus qualify for further investigation.

However, we take the problem of detecting anomalies in voter data to a new level relative to past research. First, instead of focusing our attention on a single county as in Kim, Schneider and Alvarez (2019), we use data from all sixty-seven counties in the state of Florida. This gives us the ability to not only look for outliers over instances of the dataset (*over time*), but it also gives us the ability to look for outliers across counties (*across space*). The idea is that while there can be some degree of variance between counties, we expect similar trends across them. The ability to look for statistical anomalies across time and space is one primary methodological contribution of our work.

Second, instead of considering each of the generated time-series of changes to the data separately, we use a statistical model with multiple response variables. We discuss our model in detail in the next section of the paper.

Our paper also has important substantive relevance, as we use data from the 2016 presidential general election from Florida. With repeated allegations that voter registration systems in some Florida counties in 2016 were compromised, our analysis may help to shed some light on this controversy (Sanger, 2018). We identify a small number of Florida counties where our methodology detects significant rates of change in their voter registration databases in the 2016 presidential

election. While we are not in a position to know what causes these counties to have anomalous rates of change in their voter registration systems, we argue that this application of our methodology to the Florida 2016 data demonstrates the utility of forensic analysis of critical administrative data, in particular databases that are part of American's election infrastructure.

3 Data and Methodology

For this paper, we obtained the public-release voter registration and voting history databases directly from the Florida Secretary of State. These data are provided monthly on CD-ROM.[2] For this paper, given the interest in potential voter registration database intrusion in the 2016 presidential election, we focus on these data between June 2016 and December 2016,[3] the critical final months in the presidential election cycle when the likelihood of malicious intrusion might be great (and when unintentional administrative error might be problematic). Florida's voter data management system in 2016 was described by the state as a "hybrid" system, having features of both a "bottom-up" and "top-down" system, providing for local file maintenance which would be transmitted to the state-maintained voter registration database.[4]

3.1 Matching Records in Snapshots of Voter Registration Databases

To identify the added, dropped, and changed records in snapshots of voter registration databases, we need to first match the records in consecutive snapshots. In the social sciences, record linkage is a relatively new methodology. Ansolabehere and Hersh (2017) discussed how exact matching between voter data and other sources of administrative data performs surprisingly well, and Enamorado, Fifield and Imai (2019) developed enhanced record linkage open-source software which they tested on national voter files.

In this study, We use package `voterdiffR` by Kim, Schneider and Alvarez (2019) and package `fastLink` by Enamorado, Fifield and Imai (2019) to conduct probabilistic record linkage (PRL) (Fellegi and Sunter, 1969). PRL assumes a latent match status, which indicates a match (1) or a non-match (0), for each pair of records being compared. With this, PRL models the agreement

[2] For information on how to obtain the Voter Extract Disk from the Florida Secretary of State's Office, see https://dos.myflorida.com/elections/data-statistics/voter-registration-statistics/voter-extract-disk-request/.

[3] The snapshots were generated by the Florida Division of Elections respectively on June 7th, July 12th, August 9th, September 8th, October 20th, November 1st, and December 6th of 2016.

[4] In their responses to the Election Assistance Commission's 2016 Election Administration & Voting Survey's Statutory Overview, Florida described their voter registration process in 2016 as a "hybrid" system. They wrote "As first reported to the EAC in May 2005 in response to recommended guidelines for statewide voter registration systems, Florida registration system is considered a hybrid, incorporating features of a bottom-up system and a top-down system." The rationale for the state's response was that "Information from the counties is transmitted real-time to the statewide system." (See Section B: Voter Registration, p. 10, https://www.eac.gov/research-and-data/datasets-codebooks-and-surveys).

levels (e.g. different, similar, or nearly identical) of the pair of records at each field conditional on the latent match status.

The agreement levels are measured by using exact matching for numeric values and cutoffs in Jaro–Winkler distance for character strings. PRL estimates the parameters of the observed-data likelihood function using Expectation Maximization (EM) algorithm and derive the desired match probability for each pair using Bayes rule, after assuming (1) the latent match status are independently and identically distributed; (2) given the match status, the agreement levels are independent; and (3) conditional on the latent match status, the missing values are missing at random (MAR).

By defining a single threshold S, any record pairs with match probabilities higher than S are identified as matched pairs, or non-matched pairs otherwise. Notice that these PRL identified matches/non-matches are estimates of the latent match status. We present the analysis of PRL matches in the main text, and repeat the study on another estimation of the latent match status, which uses exact-string matching, and provide the results in the Appendix C.

The fields selected to assess matching are: *last name*, *first name*, *date of birth*, *residential address (street level)*, *daytime phone number*, and *gender*. A Jaro–Winkler value of 0.88 (Jaro–Winkler value range from 0, completely different, to 1, identical) is used as the threshold in the character string matching to classify the agreement levels, and a single threshold 0.85 is used in identifying the matched and non-matched pairs.[5] After matching the snapshots[6], the added, dropped, changed records are identified as:

- **Added Records**: Records from snapshot t that have not been PRL matched with any records in snapshot $t-1$.
- **Dropped Records**: Records from snapshot $t-1$ that have not been PRL matched with any records in snapshot t.
- **Changed Records**: Records from snapshot t that have been PRL matched with existing records in snapshot $t-1$, but at least one of the fields (e.g. address, last name) are changed during the period $t-1$ to t.

Formally, we express these change rates as follows, where $c \in \{67 \text{ counties in Florida}\}$, $t \in \{Jul, Aug, Sep, Oct, No$ i is record indicator, and $k \in \{Last\ Name, Residential\ Address, Birth\ Date, Voter\ Status, Party\}$:

[5] These are the default values in `fastLink`, and tuning them do not make a substantial difference to the results.

[6] The post-merge analysis steps are outlined in Enamorado, Fifield and Imai (2019), which lays out the steps to incorporate the uncertainty in record linkage into the estimates in question. We compare the results to the case when we rely on exact matching as a benchmark, and leave the post-merge analysis for future research.

$$\widehat{Added}_{c,t} = \frac{\text{Number of Added Records (PRL) in County } c \text{ at Time } t}{\text{Average Total Records in County } c \text{ at Time } t-1 \text{ and } t}$$
$$= \frac{\sum_{i \in c} 1_{(x_{i,t} \ is \ added)}}{(N_{c,t-1} + N_{c,t})/2}$$
$$\widehat{Dropped}_{c,t} = \frac{\text{Number of Dropped Records (PRL) in County } c \text{ at Time } t-1}{\text{Average Total Records in County } c \text{ at Time } t-1 \text{ and } t}$$
$$= \frac{\sum_{i \in c} 1_{(x_{i,t-1} \ is \ dropped)}}{(N_{c,t-1} + N_{c,t})/2}$$
$$\widehat{Changed}^{k}_{c,t} = \frac{\text{Number of Changed Records (PRL) in County } c \text{ at Time } t \text{ in field } k}{\text{Average Total Records in County } c \text{ at Time } t-1 \text{ and } t}$$
$$= \frac{\sum_{i \in c} 1_{(x^k_{i,t} \ is \ changed)}}{(N_{c,t-1} + N_{c,t})/2}$$

3.2 Finding Anomalies in Voter Registration Changes

Having linked the voter records and estimated the rate of change quantities by time period for each county, we now want to identify counties that demonstrated anomalous rates of change. Changes should be expected in administrative datasets, and especially in such a contested general election. The issue is finding periods of time, and the particular counties, where the rates of change are sufficiently large as to be deemed an anomaly deserving further examination.

We want to use an approach for statistical anomaly detection that identifies counties and time periods that show anomalous rates of change in the voter registration data, without generating a large number of false positives, which could be counterproductive for election administrators. Too many false alarms that require accuracy verification will be detrimental to their work, and it could also unnecessarily erode stakeholder and voter confidence in the integrity of the data and the election.

Thus, in this paper, we use two complementary approaches for identifying Florida counties with potentially anomalous rates of change in the voter registration data that we use, similar to the techniques used in the more general literature on statistical anomaly detection (Chandola, Banerjee and Kumar, 2009). First, we use a simple visual presentation of box-and-whisker plots to compare the univariate distribution of change statistics across Florida counties. Second, we provide a Bayesian multivariate analysis to model the change rates and check for counties that have patterns of record changes that deviate from other counties. It is often the case that anomalies apparent with simple visualizations are confirmed with more sophisticated multivariate analysis. In the next subsection, we first explain the second method.

3.3 A Bayesian Approach

For a principled detection of anomalies that may have occurred in snapshots of the Florida voter registration database, we conduct a Bayesian analysis, which first estimates a model that best explains the variance within changes to the data (e.g. added rates, dropped rates, and changed rates), then creates predictions from the posterior parameter distributions, and finally, identifies deviations that have significantly large residuals.

With the assumption that the data inconsistencies within and across jurisdictions are correlated, instead of using multiple univariate models, we use a multivariate model to gain further insight into the anomalies. While the terminology 'multivariate' is usually reserved to indicate a regression model containing multiple covariates, in this case we mean a model with *multiple response variables*. This use of a multivariate model to detect anomalies in voter list maintenance is one of the primary contributions of this paper.

The multivariate models can be estimated using frequentist methods such as maximum likelihood, but as frequentist approaches cannot easily incorporate prior knowledge about the data generation process, they are not as useful as Bayesian approaches for estimating complex models. Instead of treating the unknown parameters as fixed constants, Bayesian methods assume that they are random variables. Starting from the prior information, Bayesian methods recursively derive inferences in a consistent way from the data and push the posterior distributions of the parameters close to the sample information. In this study, we use the package `brms` (Bürkner et al., 2017) and the package `stan` (Stan Development Team, 2019) to conduct Hamiltonian Monte Carlo (Duane et al., 1987). With the help of No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), HMC converges much faster than other Markov Chain Monte Carlo (MCMC) methods such as Metropolis-Hastings updating (Chib and Greenberg, 1995; Hastings, 1970) and Gibbs-sampling (Damlen, Wakefield and Walker, 1999; Neal et al., 2011).

Our Bayesian multivariate analysis follows the steps:

- **Step 1**: Use a multivariate model with changes to the voter data (e.g. added rates, dropped rates, and change rates) as dependent variables and model county-level and time-level random effects. We also include county-level heterogeneity such as population, which could affect the rates of change—especially in smaller counties where the proportion of changes could be inflated due to high variance.
- **Step 2**: Estimate the multivariate model using Hamiltonian Monte Carlo (HMC), and obtain the converged distributions of parameters.
- **Step 3**: Predict the dependent variables by making 24,000 random draws from the posteriors, and then computing the mean of each set of draws.
- **Step 4**: Quantify the differences between the dependent variables and the model predictions by using t-scores from the hypothesis tests with the null hypothesis that *residual is zero*.

· **Step 5**: Identify the significant deviations that have t-scores larger than 1.96.

The multivariate model can be written as follows:

$$Y_{n\times p} = h(X_{n\times(r+1)}\beta_{(r+1)\times p}) + \epsilon_{n\times p}$$

where $Y_{n\times p}$ is a list of dependent variables:

$$Y_{n\times p} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_p \end{bmatrix},$$

Each row of $Y_{n\times p}$ stands for a *county*$\otimes$*time* combination, so that $n = n_{county} \times n_{time} = 67 \times 6 = 402$. Columns of $Y_{n\times p}$, i.e., $[\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_p]$ are $p = 7$ types of PRL estimated change rates, which are {*Added Rates*, *Dropped Rates*, *Changed Rates in Last Name, Residential Address, Birth Date, Voter Status, and Party*}. $X_{n\times(r+1)}$ is a list of explanatory variables:

$$X_{n\times(r+1)} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{1,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{2,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n_{\text{county}},1} & x_{n_{\text{time}},2} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix},$$

Here, $[\mathbf{x}_1 \quad \mathbf{x}_2]$ are group level variables {*County*, *Time*}. For the rest of the model, $\beta_{(r+1)\times p}$ are coefficients and $\epsilon_{n\times p}$ is error structure:

$$\epsilon_{n\times p} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \cdots & \epsilon_{1,p} \\ \epsilon_{2,1} & \epsilon_{2,2} & \cdots & \epsilon_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \cdots & \epsilon_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1' \\ \boldsymbol{\epsilon}_2' \\ \vdots \\ \boldsymbol{\epsilon}_n' \end{bmatrix}$$

The assumption is that $\mathrm{E}(\epsilon_i) = 0$ for all $i = 1, \cdots, p$ and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = \sigma_{ij} I$ for all $i, j = 1, \cdots, p$.

$h(\cdot)$ is Bayesian estimation that assumes dependent variables follow zero-inflated beta (ZIB) distributions. We use ZIB to harness the flexibility of the Beta distribution to model proportion data, while also accounting for the fact that the data can contain a substantial amount of zeros (Ospina

and Ferrari, 2012):

$$y_{i,j} \sim ZIB(f(x_i\beta_j), \theta_{i,j})$$

where i denotes the observation and j denotes the dependent variable. $\theta_{i,j}$ includes three parameters that specify a ZIB distribution, $\{\alpha_{i,j}, \mu_{i,j}, \phi_{i,j}\}$. The density function of ZIB distribution can be written as:

$$ZIB(y_{i,j}; \alpha_{i,j}, \mu_{i,j}, \phi_{i,j}) = \begin{cases} \alpha_{i,j} & \text{if} \quad y_{i,j} = 0 \\ (1-\alpha_{i,j})B(y_{i,j}; \mu_{i,j}, \phi_{i,j}) & \text{if} \quad 0 < y_{i,j} < 1 \end{cases}$$

$B(\cdot)$ is Beta density function (Ferrari and Cribari-Neto, 2004):

$$B(y_{i,j}; \mu_{i,j}, \phi_{i,j}) = \frac{\Gamma(\phi_{i,j})}{\Gamma(\mu_{i,j}\phi_{i,j})\Gamma((1-\mu_{i,j})\phi_{i,j})} y_{i,j}^{\mu_{i,j}\phi_{i,j}-1}(1-y_{i,j})^{(1-\mu_{i,j})\phi_{i,j}-1}$$

$\Gamma(\cdot)$ is the gamma function with the following parameterization:

$$\begin{cases} \mathrm{E}(y_{i,j}) &= \mu_{i,j} \\ \mathrm{Var}(y_{i,j}) &= \frac{\mu_{i,j}(1-\mu_{i,j})}{\phi_{i,j}+1} \end{cases}$$

In this study, we use the logit link function[7] for estimates of $\mu_{i,j}$, which is

$$\ln \frac{\Pr(\mu_{i,j} = 1 \mid x_i)}{1 - \Pr(\mu_{i,j} = 1 \mid x_i)} = x_i\beta j,$$

and use identity link functions for $\alpha_{i,j}$ and $\phi_{i,j}$.

Once we have run our Hamiltonian Monte Carlo (Duane et al., 1987; Neal et al., 2011) estimation, we use the fitted model to make predictions for rates of change that we can compare to the actual observed data. This yields a simple and straightforward test statistic, as we note that the $y_{c,t}$ which deviates from the model prediction can be found using the following hypothesis test:

$$H_0 : \ y_{i,j} = \hat{y}_{i,j}$$
$$H_a : \ y_{i,j} \neq \hat{y}_{i,j}$$

[7]Other link functions such as probit produce estimates of μ that are highly equivalent (Gunduz and Fokoué, 2015). We show the results of Bayesian multivariate analysis with a Cauchit link function in the Appendix E.

4 Results

4.1 Visual Comparison Across Florida Counties

Our first approach for finding outliers is to visualize the distributions of these change statistics. We use box-and-whisker plots, as they provide a concise representation of distributions and allow for easy visualization of outliers. The boxes indicates the first and third quartiles (lower and upper hinges), while the whiskers indicate the smallest to largest original data point at most 1.5 × the interquartile range (IQR). Box-and-whisker plots are in Figures 1 to 2, where the change rates are grouped by time. While the 1.5 × IQR detects the mild anomalies, we labeled only the extreme outliers (outside 3 × IQR fences) with their respective county name abbreviations. See the Appendix A for a crosswalk between abbreviations and full county names.

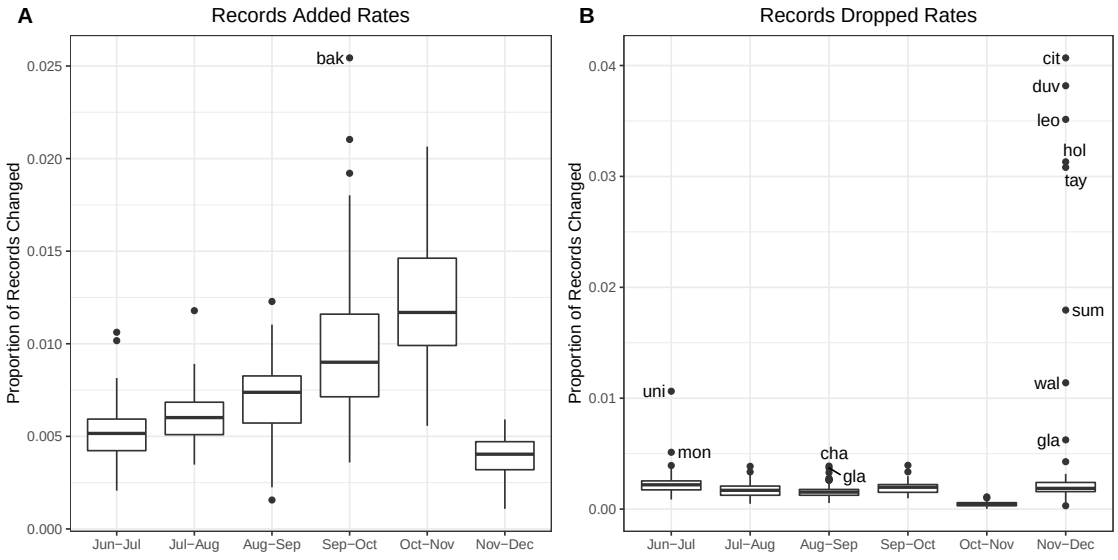

Figure 1: Proportion of Records Added and Dropped

The first interesting observation is that change rates are very different per metric. The proportion of records added generally increase throughout the pre-election period while falling sharply after the election. The proportion of records dropped show generally a downward trend, with extreme outliers in November to December. The proportion of records changed in the five fields are generally similar between June to September, sharply climbing in September to October, and then dropping in October to November. Note that the general election date was November 8 in 2016, and therefore the 29-day registration deadline was at October 18, 2016.

Given these common local trends, we note 69 outliers outside the 3 × IQR range from Figures 1–2. The most extreme outliers we see are in the proportion of records dropped (Figure 1, panel B): from November to December 2016, seven counties had very high record-dropping rates: Citrus, Duval, Leon, Holmes, Taylor, Sumter, and Walton, which are scattered beyond even the 5 × IQR

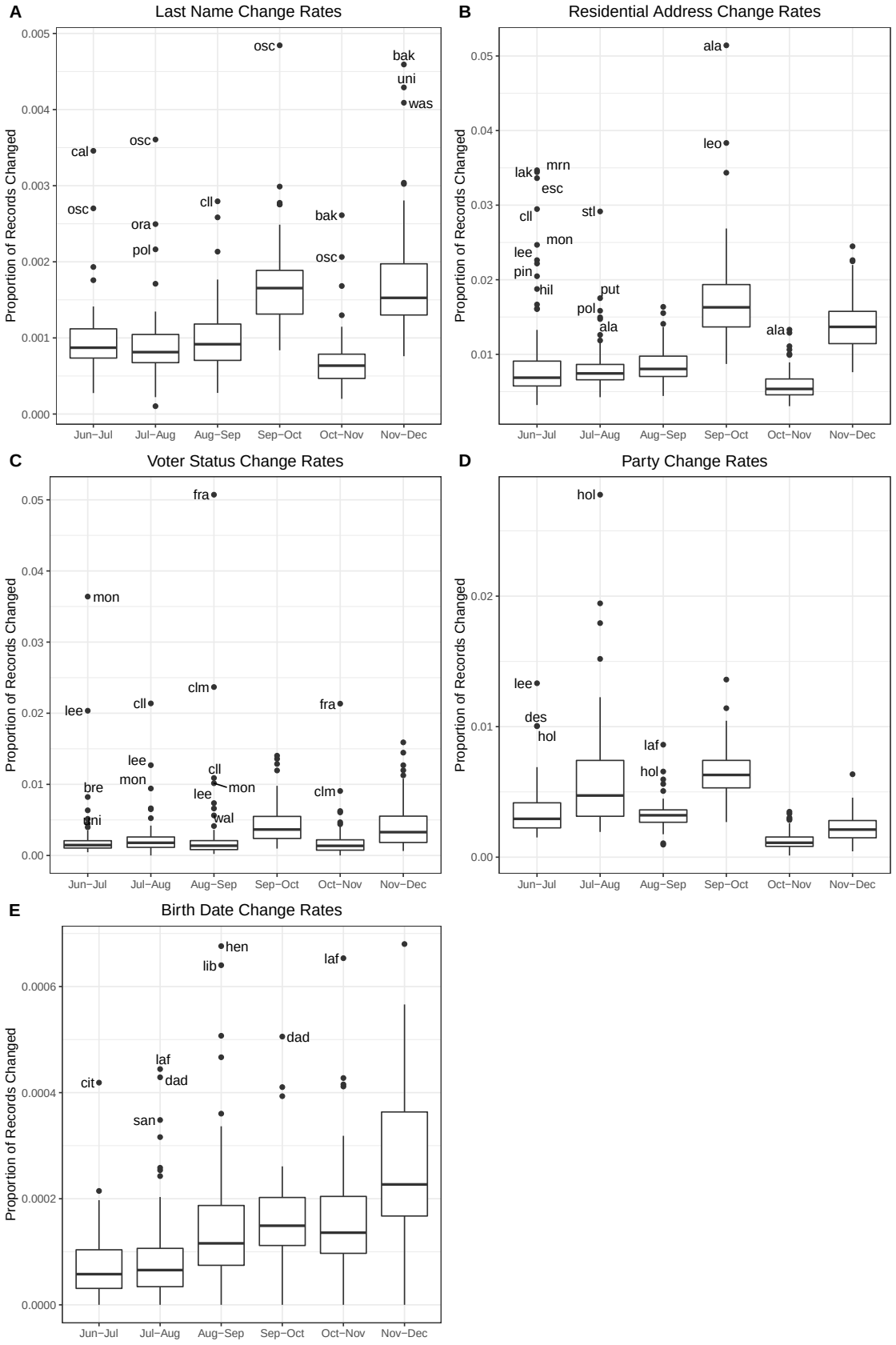

Figure 2: Proportion of Records Changed by Field

limit. These are much more pronounced than other plotted anomalies in either Figure 1, panel A, or Figure 2.

These are not necessarily evidence of deliberate internal or external data manipulation in these counties. Since the Division of Elections in Florida exported the November 2016 snapshot on the first day of the month, these are likely maintenance activities after the general election date—any data manipulation aimed at influencing the election would have be done before the election. These results simply indicate that given the sixty-seven counties in Florida, rates of dropped voters in certain counties are significantly different compared to the rest of the counties' distributions per month.

Setting aside the most extreme outliers, we can explore the outliers using various different criteria. For instance, we may be interested in counties that repeatedly emerge as anomalies for the same metric, such as Osceola County, which had high last name change rates in four time periods—Jun-Jul, Jul-Aug, Sep-Oct, and Oct-Nov. Given that it is one of the bigger counties (2016 population estimate of 322,862), it might be worth further investigation as to why Osceola County had higher rates of last name changes. Or take Monroe or Lee Counties, which both appear to have high voter status change rates from Jun-Jul, Jul-Aug, and Aug-Sep. Given that the voter status change means a switch between active, inactive, to preregistered, we may take more interest in these counties. What to investigate depends on the domain expertise of what may constitute threats to the accuracy and the integrity of the data.

While visualizations have offered an initial overview of the utility of our methodology, for a principled investigation we need a statistical model to pool the information across time, while also incorporating county-level heterogeneity. In particular, as aforementioned, we intend to build a model with multiple response variables to account for certain common patterns in the response variables.

4.2 Identifying Deviations from Multivariate Model Prediction

Moving from the graphical presentation, we now present our multivariate Bayesian model to capture the dynamics between change rates, populations, and group-level variables such as county and time, and then use model prediction to detect deviations. The t scores of the hypothesis tests are shown in Figures 3 to 4. t score higher than 1.96 indicates $y_{c,t}$ is significantly ($\alpha < 0.05$) different from model prediction $\hat{y}_{c,t}$.

There were many "anomalies"—79 in total, which constitute almost 20% of the data—in Figures 3 to 4. Note that the labelled anomalies are quite similar to what lay outside the 3 $\times$ IQR fence in Figures 1 and 2.

The following anomalies are significantly ($\alpha < 0.001$) different from the model prediction:

1. From Nov-Dec, as previously noted in the box-and-whisker plots, Citrus, Duval, Leon, Holmes, Taylor, Sumter, and Walton County had drop rates significantly ($\alpha < 0.00001$) different from

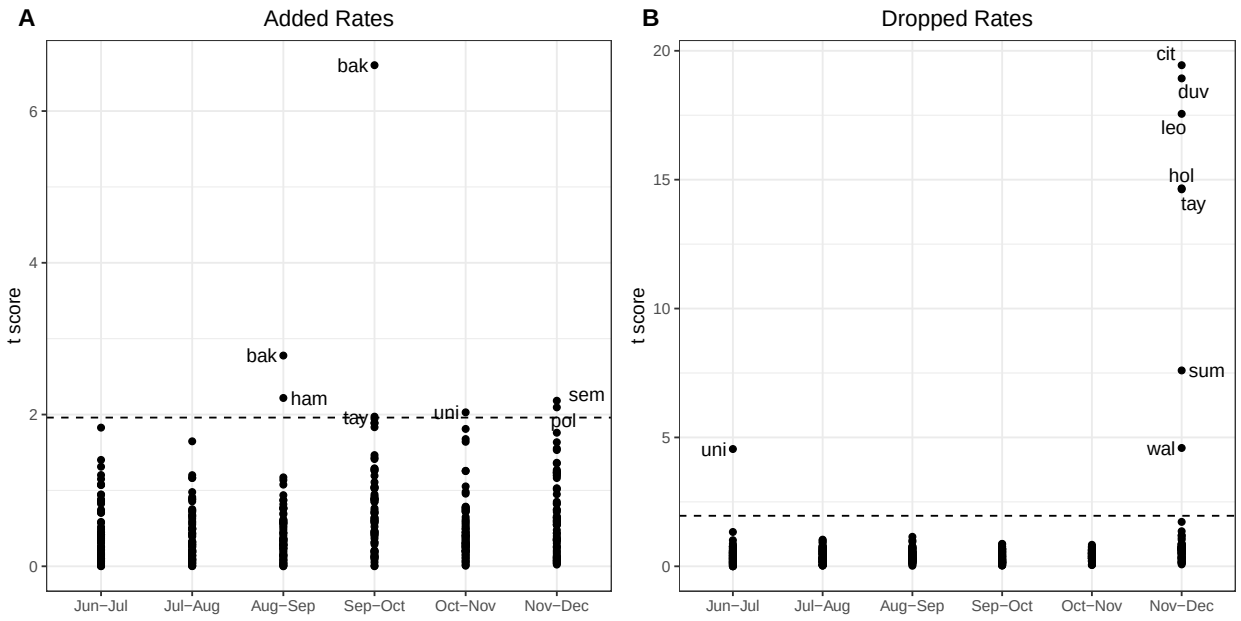

Figure 3: Deviations of Added and Dropped Rates

model prediction.

2. Baker County had significantly ($\alpha < 0.00001$) different add rate in Aug-Sep and Sep-Oct, and significantly ($\alpha < 0.0005$) different last name change rate in Oct-Nov and Nov-Dec.
3. Calhoun County had significantly ($\alpha < 0.00001$) different last name change rate in Jun-Jul, and significantly ($\alpha < 0.00001$) different party change rate in Jul-Aug.
4. Collier County had significantly ($\alpha < 0.0005$) different last name change rate in Aug-Sep, significantly ($\alpha < 0.00001$) different residential address change rate in Aug-Sep, and significantly ($\alpha < 0.0001$) different voter status change rate in Jul-Aug.
5. Marion, Lake, Escambia, and Pinellas County had significantly ($\alpha < 0.00001$) different residential address change rates in Jun-Jul.
6. St. Lucie County had significantly ($\alpha < 0.00001$) different residential address change rate in Jul-Aug.
7. Alachua County had significantly ($\alpha < 0.0005$) different residential address change rate in Sep-Oct.
8. Monroe and Brevard County had significantly ($\alpha < 0.0005$) different voter status change rate in Jun-Jul.
9. Franklin County had significantly ($\alpha < 0.00001$) different voter status change rate in Aug-Sep and Oct-Nov.
10. Columbia County had significantly ($\alpha < 0.00001$) different voter status change rate in Aug-Sep.
11. Lee County had significantly ($\alpha < 0.0005$) different voter status change rate in Jun-Jul, and significantly ($\alpha < 0.00001$) different party change rate in Jun-Jul.

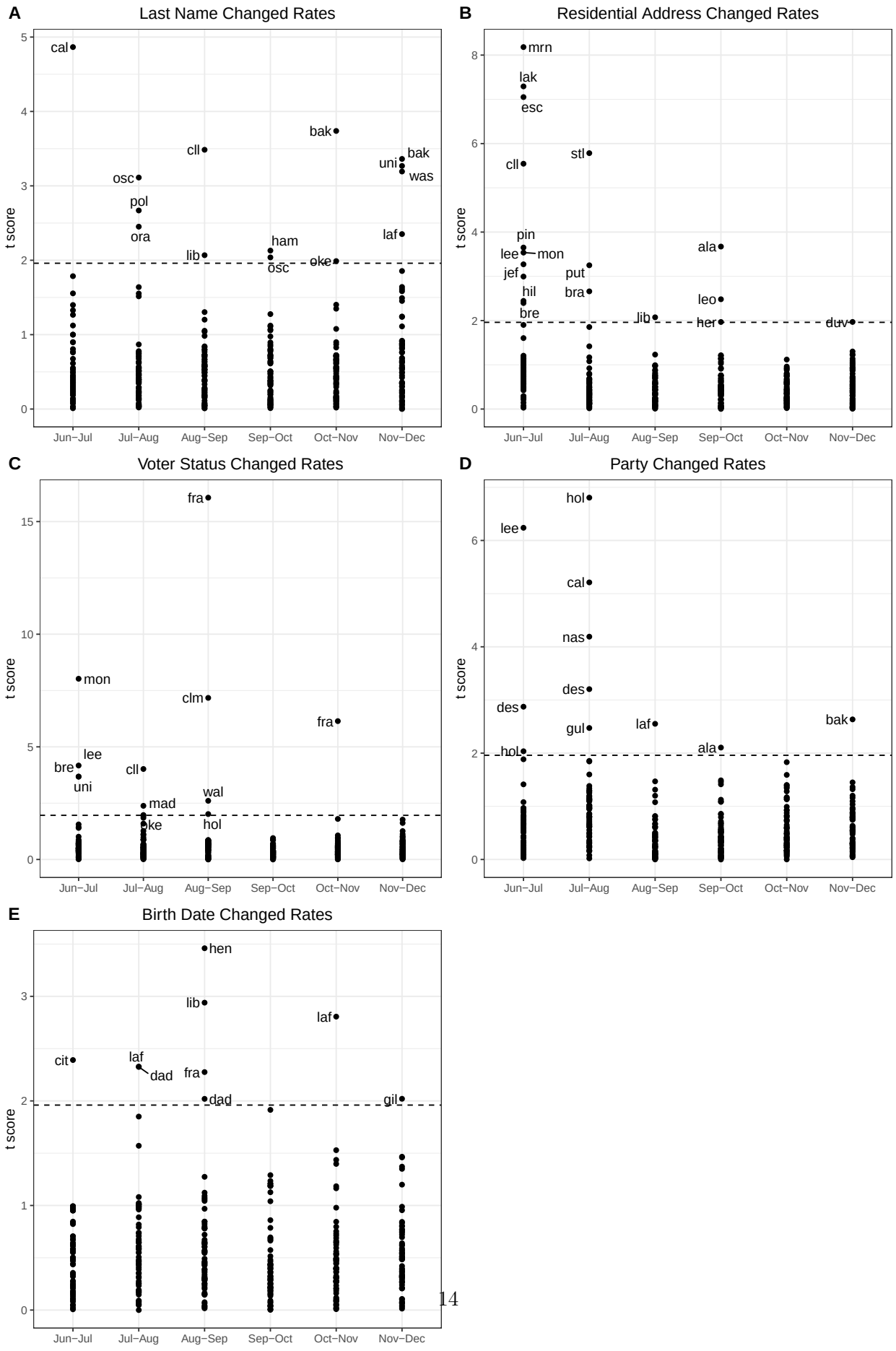

Figure 4: Deviations of Changed Rates

12. Holmes and Nassau County had significantly ($\alpha < 0.0001$) different party change rates in Jul-Aug.
13. Hendry County had significantly ($\alpha < 0.0005$) different birth date change rate in Aug-Sep.

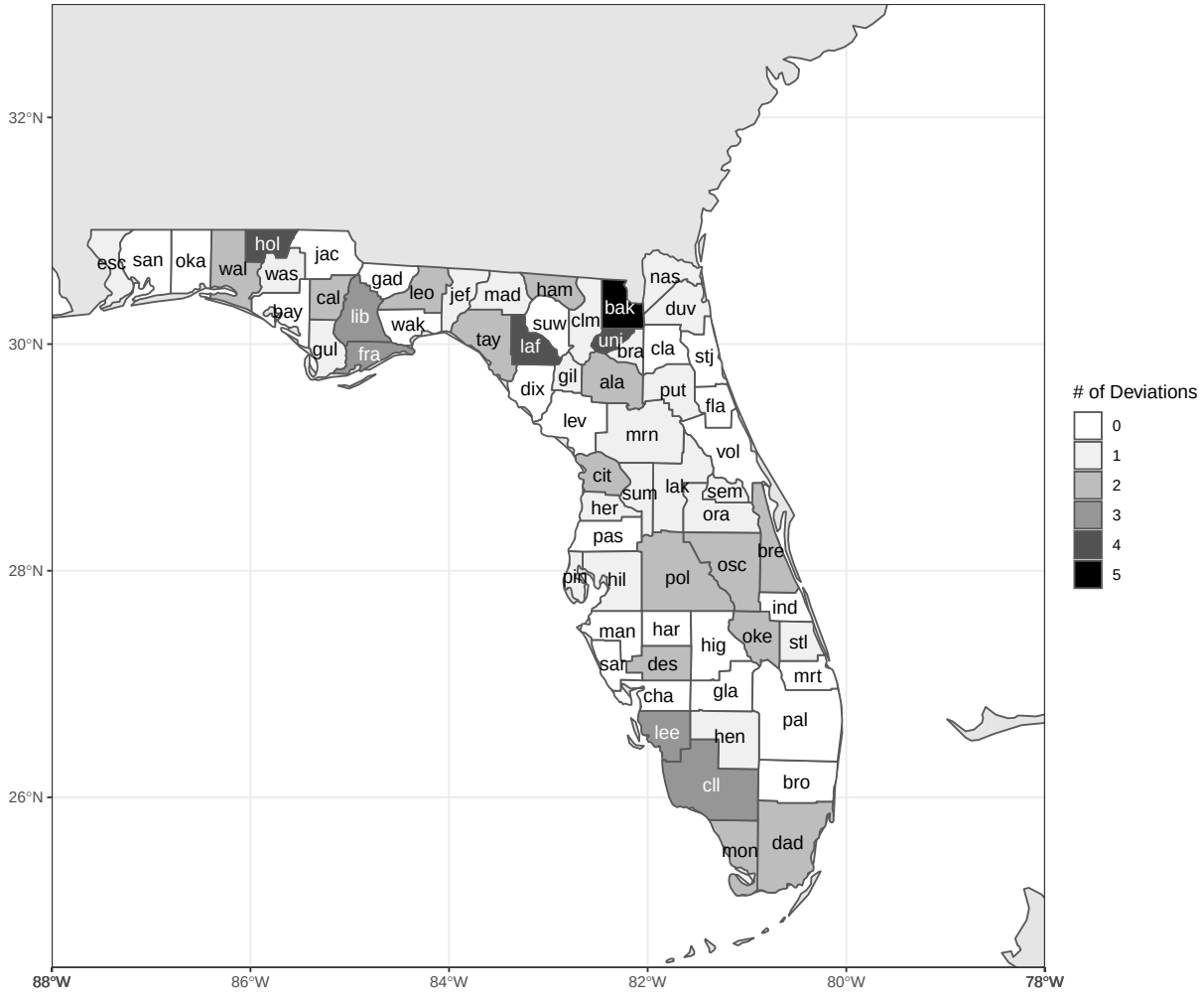

Figure 5: Number of Deviations from the Model Per County

As can be seen, there are many anomalies listed, which is difficult to digest. To aggregate these deviations into a more succinct picture, Figure 5 shows, out of 42 events (7 metrics × 6 snapshots), how many times a county deviated from the model prediction. There were 42 counties that had at least one deviation and 22 counties that had at least two deviations. Collier, Franklin, Lee, and Liberty County had three deviations. Holmes, Lafayette, and Union had four. Baker County, which

was the one with the most deviations, had its change rates significantly ($\alpha < 0.05$) different from model predictions in 5 events. If our interest lies in seeing which counties have very different voter list maintenance results, we can now sort the counties by these order, and Figure 5 can serve as a benchmark.

So how exactly do list maintenance practices of Baker, Holmes, Lafayette, and Union counties differ from the rest of the states? Again, this is difficult to answer, and we want to emphasize that heterogeneity in voter list maintenance practices can come from both internal and external factors, and does not necessarily imply deliberate internal or external manipulation of the voter registration data. Indeed, there's very little transparency about list maintenance in individual counties, in particular in smaller-population counties. Indeed, by Florida counties' 2016 population estimates (The Office of Economic and Demographic Research, 2020), Baker had 26,965 countywide residents, Holmes 20,003, Lafayette 8,621, and Union 15,887—they are all small counties.

But this shows that our model can serve as the first step to check the accuracy and integrity of a statewide voter database, that depends on the aggregation of sub-state databases, which potentially are products of heterogeneous voter maintenance practices. Figure 5 shows one way to summarize and look for the most pronounced anomalies. Aggregation by metric or just one metric of interest such as voter status change rates—as aforementioned—are other ways to summarize these detected anomalies and make sense from them.

Setting aside potentially different list maintenance practices, note that there are two Florida counties that have been reported by the media to have had its data breaches, and among them one is Washington County. Its data does not seem to display any particularly deviant patterns according to our model. This seems consistent with the official report by the FBI and the Department of Homeland Security that there was no evidence of data alterations, although the affected counties' systems had been breached. We have yet to learn from authorities which other county had the data breach. Lastly, we perform the same analysis while accounting for county-level partisan composition (percentage of registrants who are Republican)—the results are qualitatively similar to what we report here (see Appendix D).

5 Discussion and Conclusion

How do we ensure the accuracy and integrity of a statewide voter registration data, given that in many states it is the aggregation of local jurisdictions' data, or the result of a combination of state and local list maintenance practices? Voter list maintenance procedures potentially vary by county, and we would like to compare the result of these decentralized practices. In this paper, we provided a benchmark model to find anomalies in county-level rates of changes to the voter data that accounts for common patterns in local data and local-level heterogeneity. Using Florida's monthly data in the 2016 cycle as an example, we indeed find a great deal of heterogeneity in counties' voter list

maintenance results, even within the same state.

We first used probabilistic record linkage to estimate the proportion of changes to the voter file using seven metrics—voters added, voters deleted, last name changed, address changed, birth date changed, voter status changed, and party changed. We then used visualizations and a Bayesian multivariate multilevel model to identify counties that have change rates that significantly differ from others. Major findings are robust to different matching schemes, different link functions in the Bayesian model, or adding another county-level covariate.

Baker, Holmes, Lafayette, and Union counties particularly stood out in their number of deviations from the model prediction. In addition, Citrus, Duval, Leon, Holmes, Taylor, Sumter, and Walton counties had high proportions of voters dropped from November to December. The anomalies that were pronounced from the model did not include Washington county, which was allegedly one of the two counties whose data was breached by foreign intelligence in 2016.

Again, the results do not indicate that these counties' list maintenance practices are necessarily problematic—they are simply very different compared to other counties, controlling for various things. But as can be seen, our results serve as a benchmark to begin looking for potential deviance, which could threaten accuracy and integrity of statewide voter databases.

In conclusion, voter registration databases form the foundation for many election administration activities in the United States. These databases are used to allocate resources for early and Election Day voting. They are used to authenticate voters to obtain and cast their ballots, whether in-person or by-mail. The data are used by election officials to send voter information materials, and by political campaigns and other organizations for canvassing and get-out-the-vote activities. The results of studies like ours can be used to investigate the observed anomalies, and to thus improve both list maintenance practices and eventually improve the accuracy and integrity of these important administrative databases.

References

Alvarez, R. Michael. 2005. "Potential Threats to Statewide Voter Registration Systems.".

Alvarez, R. Michael, Jeff Jonas, William E. Winkler and Rebecca N. Wright. 2009. "Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project." *Proceedings of the 2009 Electronic Voting Technology Workshop-Workshop on Trustworthy Elections* .
URL: *https://www.usenix.org/legacy/events/evtwote09/tech/full_papers/alvarez.pdf*

Ansolabehere, Stephen and Eitan Hersh. 2010. "The quality of voter registration records: A state-by-state analysis." *Report, Caltech/MIT Voting Technology Project* .

Ansolabehere, Stephen and Eitan Hersh. 2014. "Voter registration: the process and quality of lists." *The measure of American elections* pp. 61–90.

Ansolabehere, Stephen and Eitan Hersh. 2017. "ADGN: an algorithm for record linkage using address, date of birth, gender, and name." *Statistics and Public Policy* 4(1):1–10.

Berent, Matthew K., Jon A. Krosnick and Arthur Lupia. 2016. "Measuring Voter Registration and Turnout in Surveys: Do Official Government Records Yield More Accurate Assessments?" *Public Opinion Quarterly* 80(3):597–621.

Bürkner, Paul-Christian et al. 2017. "brms: An R package for Bayesian multilevel models using Stan." *Journal of Statistical Software* 80(1):1–28.

Caltech/MIT Voting Technology Project. 2001. "Voting: What Is, What Could Be.".
URL: *vote.caltech.edu/documents/153/voting_what_is_what_could_be*

Chandola, Varun, Arindam Banerjee and Vipin Kumar. 2009. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41(3):15.

Chib, Siddhartha and Edward Greenberg. 1995. "Understanding the metropolis-hastings algorithm." *The american statistician* 49(4):327–335.

Christen, Peter. 2012. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Christen, Peter. 2014. "Preparation of a real voter data set for record linkage and duplicate detection research.".

Christen, Peter and Ross W. Gayler. 2013. "Adaptive Temporary Entity Resolution on Dynamic Databases." *Advances in Knowledge Discovery and Data Mining, PAKDD 2013, edited by J. Pei, V.S. Tseng, L. Cao, H. Motoda, and G. Xu, Lecture Notes in Computer Science, Springer* 7819:558–569.

Damlen, P, John Wakefield and Stephen Walker. 1999. "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(2):331–344.

Duane, Simon, Anthony D Kennedy, Brian J Pendleton and Duncan Roweth. 1987. "Hybrid monte carlo." *Physics letters B* 195(2):216–222.

Enamorado, Ted, Benjamin Fifield and Kosuke Imai. 2019. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." *American Political Science Review* 113(2):353–371.

Fellegi, Ivan P and Alan B Sunter. 1969. "A theory for record linkage." *Journal of the American Statistical Association* 64(328):1183–1210.

Ferrari, Silvia and Francisco Cribari-Neto. 2004. "Beta regression for modelling rates and proportions." *Journal of applied statistics* 31(7):799–815.

Green, Donald P. and Alan S. Gerber. 2006. "Can Registration-Based Sampling Improve the Accuracy of Midterm Election Forecasts?" *The Public Opinion Quarterly* 70(2):197–223.

Gunduz, Necla and Ernest Fokoué. 2015. "On the predictive properties of binary link functions." *arXiv preprint arXiv:1502.04742* .

Hastings, W Keith. 1970. "Monte Carlo sampling methods using Markov chains and their applications.".

Hoffman, Matthew D and Andrew Gelman. 2014. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15(1):1593–1623.

Kim, Seo-young Silvia, Spencer Schneider and R. Michael Alvarez. 2019. "Evaluating the Quality of Changes in Voter Registration Databases." *American Politics Research* .

Liebschutz, Sarah F. and Daniel J. Palazzolo. 2005. "HAVA and the States." *Publius: The Journal of Federalism* 35:497–514.

Merivaki, Thessalia. 2019. "Access Denied? Investigating Voter Registration Rejections in Florida." *State Politics & Policy Quarterly* 19(1):53–82.
URL: *http://journals.sagepub.com/doi/10.1177/1532440018800334*

Merivaki, Thessalia. 2020. ""Our Voter Rolls Are Cleaner Than Yours": Balancing Access and Integrity in Voter List Maintenance." *American Politics Research* p. 1532673X2090647.
URL: *http://journals.sagepub.com/doi/10.1177/1532673X20906472*

Mueller, Robert S. 2019. Report on the investigation into Russian interference in the 2016 presidential election. Technical report US Dept. of Justice. Washington, DC.

Neal, Radford M et al. 2011. "MCMC using Hamiltonian dynamics." *Handbook of markov chain monte carlo* 2(11):2.

Ospina, Raydonal and Silvia LP Ferrari. 2012. "A general class of zero-or-one inflated beta regression models." *Computational Statistics & Data Analysis* 56(6):1609–1623.

Pettigrew, Stephen and Charles Stewart III. 2017. "Moved Out, Moved On: Assessing the Effectiveness of Voter Registration List Maintenance." *MIT Political Science Department Research Paper No. 2018-1* .

Sanger, David E. 2018. *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age*. Crown.

Stan Development Team. 2019. "RStan: the R interface to Stan.". R package version 2.19.2.
URL: *http://mc-stan.org/*

The Office of Economic and Demographic Research. 2020. "Population and Demographic Data - Florida Products.".
URL: *http://edr.state.fl.us/Content/population-demographics/data/index-floridaproducts.cfm*