Article

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

Bayesian Analysis of State Voter Registration Database Integrity

https://doi.org/10.1515/spp-2021-0016
Received July 6, 2021; accepted December 13, 2021

Abstract: How do we ensure a statewide voter registration database's accuracy and integrity, especially when the database depends on aggregating decentralized, sub-state data with different list maintenance practices? We develop a Bayesian multivariate multilevel model to account for correlated patterns of change over time in multiple response variables, and label statewide anomalies using deviations from model predictions. We apply our model to California's 22 million registered voters, using 25 snapshots from the 2020 presidential election. We estimate countywide change rates for multiple response variables such as changes in voter's partisan affiliation and jointly model these changes. The model outperforms a simple interquartile range (IQR) detection when tested with synthetic data. This is a proof-of-concept that demonstrates the utility of the Bayesian methodology, as despite the heterogeneity in list maintenance practices, a principled, statistical approach is useful. At the county level, the total numbers of anomalies are positively correlated with the average election cost per registered voter between 2017 and 2019. Given the recent efforts to modernize and secure voter list maintenance procedures in the *For the People Act of 2021*, we argue that checking whether counties or municipalities are behaving similarly at the state level is also an essential step in ensuring electoral integrity.

Keywords: Bayesian anomaly detection, file linkage, administrative data, voter registration

***Corresponding author: Jian Cao**, California Institute of Technology, Pasadena, CA, USA, E-mail: jccit@caltech.edu
Seo-young Silvia Kim, American University, Washington, DC, USA. https://orcid.org/0000-0002-8801-9210
R. Michael Alvarez, California Institute of Technology, Pasadena, CA, USA. https://orcid.org/0000-0002-8113-4451

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

Twenty years later, the United States has gone through another very close election, and there are new calls for election reform—particularly the recent "*For the People Act of 2021*". This is an all-encompassing bill aimed at addressing "voter access, election integrity and security, campaign finance, and ethics for the three branches of government" (117th Congress [2021–2022], 2021). It is indeed high time for a sweeping reform—and accompanying federal funding—in election administration infrastructure for reliability, accuracy, integrity, and security of elections. Many states' voter registration systems are outdated, and needing to be updated or replaced.

However, updating and replacing election administration systems does not automatically ensure a better electoral process. We argue that monitoring and auditing of statewide voter databases, which depend on aggregating decentralized, sub-state data with varying list maintenance practices, is another essential but neglected aspect of election administration. The statewide, electronic voter registration databases mandated by HAVA have undergone little external analysis. Thus the research community, stakeholders, and the public knows little about the accuracy and security of these critical databases. We present a model that aims to accomplish this goal, which can determine whether these countywide or municipal-level databases are changing consistently within a given state or whether they are generating anomalous patterns needing additional investigation.

Why is this important? Election administration in American politics is highly decentralized, unlike the rest of the world (Bowler et al. 2015). American states conduct elections with little direction from the federal government. But further decentralization in election administration is typical in the states, as sub-state

1 The *Help American Vote Act* is P.L. 107–252 (2002). The complete text is available at https://www.eac.gov/assets/1/6/HAVA41.pdf.

jurisdictions (counties or municipalities in some states) have unique management practices, including the details about how they maintain their voter databases. In fact, the administrative complexity varies greatly across sub-state jurisdictions (Burden et al. 2012): much depends on the size and diversity of the jurisdiction's population, their budget and staffing, and the choice of private technology providers. One might imagine that some oversight mechanisms might exist to flag if some of these practices in a jurisdiction produced administrative outcomes that wildly differed from other outcomes in the rest of the state. However, there typically are not such early warning systems in place, particularly for stakeholders (like political parties and candidates who use these data and must trust their integrity when these data are used in an election), researchers, and the broader electorate.

Voter registration data provides the foundation for many of the critical components of election administration, and thus having accurate registration data is important. Voter registration data is used before elections to determine how to allocate polling places and voting centers, staff those voting locations, and resource them for each election. Registration data is also used to send voter information guides and by campaigns to mobilize their supporters. Accurate voter registration data is necessary for efficient election administration—it will reduce the number of mail ballots that might be sent to incorrect addresses, resolve those who changed residential addresses, and remove deceased voters. More accurate data will also mean fewer provisional ballots are used for in-person voting, which reduces lines in voting locations and lessens the post-election need to verify each provisional ballot.

Maintenance of voter data is extraordinarily complex and demanding work, and sometimes things may go awry even with the best intentions. In California, it was reported that automatic voter registration produced duplicate records and incorrect partisan registration in 2018, and in some counties, voters may have been incorrectly removed from voter lists used for in-person voting.[2] In Ohio, amid controversies of voter purges, differential purge rates were reported within the state, and Franklin County, in particular, was under fire for wrongly canceling voter registrations (Rouan 2020). List maintenance issues like these might be considered internal threats to a voter registration database's integrity. In fact, *The Columbus Dispatch*, the local newspaper covering these issues, stated that "Private election vendors face little oversight in Ohio despite ballot, voter purge errors," and noted that the underlying reason was administrative decentralization to local election boards.

2 See Kim, Schneider, and Alvarez 2019 for discussion of these issues in California.

To the extent to which varying local voter list maintenance practices lead to differences in the quality and accuracy of voter data, this could lead to voting rights concerns. For example, consider a county where a state agency serving that county's residents uses a malfunctioning system that incorrectly changes the addresses of registered voters. Unless these changes are quickly identified, affected registered voters may not receive voting information in the mail, they may not receive vote-by-mail ballots, and if they show up and try to vote in person they may be required to cast a provisional ballot. If these issues do not affect registered voters in other counties in the state, this is a potential issue of equal access.

Furthermore, in 2016 and 2018, media reports alleged that hackers might have tried to access state voter registration and election administration systems across the United States. These reports 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 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 list maintenance practices produce accurate and reliable data for election administration (protecting against internal threats) and that these databases are secure (protecting against external threats).

Our approach uses a Bayesian multivariate multilevel model, and can confirm that changes to the database at the sub-state level are consistent across local jurisdictions. Accounting for similarities and differences between distinctly managed sub-state databases is an essential step in assessing internal and external problems below the state level. For this, we estimate countywide change rates for multiple response variables, such as changes in voter's partisan affiliation or changes in residential addresses. We then jointly model these changes so that the changes across multiple dependent variables are correlated and label statewide anomalies using deviations from model predictions. Consistent with the literature, we show that there is a high degree of local heterogeneity in list maintenance.

We show that when the model is evaluated with synthetic data containing artificially created anomalies, its diagnostics perform better than a simple interquartile range (IQR) detection. Not all anomalies detected via this model should automatically be assumed to be problematic. Indeed, this model would require state-by-state customization to account for different practices at the state level. However, variations of this model will provide an opportunity to see, from the state's point of view, (1) what the underlying reason is for a particular jurisdiction to have high deviation from other jurisdictions, and (2) what the state can do to help resolve the diagnosed situation, if something is indeed problematic. Overall, given the recent efforts to modernize and secure voter list maintenance procedures in election reform legislation like the *For the People Act of 2021*, we argue that

checking whether counties or municipalities behave similarly at the state level is also an essential step in ensuring electoral integrity.

2 Voter Registration Database Integrity

Concerns about the integrity of voter registration data have been raised in the media and public discussion. But there has been insufficient academic research on how to measure the quality of voter registration data, detect errors, and scan for evidence of possible intrusion into these administrative databases. While some academics raised early concerns about the integrity and security of these large administrative databases after the 2000 presidential election (Alvarez 2005; Caltech/MIT Voting Technology Project 2001), it took nearly a decade for researchers to begin building methodologies for confirming the integrity and accuracy of voter registration data.

Earlier research used descriptive methods for validating voter datasets. An early effort linked voter registration data between adjacent states of Oregon and Washington 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 their information—to examine the accuracy of the 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 Arthur 2016; Green and Gerber 2006).

More recently, researchers have begun to dig 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) 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). Shino et al. (2020) shows that almost 18% of registrants report that their personal information in the voter file is inaccurate. Goel et al. (2020), while estimating the prevalence of double voting, digs deeply into potential measurement errors in voter files such as the distribution of birth dates or duplicates.

Recent research has started to focus on how different list maintenance practices at the *local* level could create heterogeneity in *statewide* voter registration data quality, particularly in states with bottom-up or hybrid list maintenance practices. Merivaki (2019) investigated how rejections of voter registration applications vary locally, depending on the time of the year as well as the source of the registration application. Also, Merivaki (2020) investigated how local registration and voting

history errors are dependent on local socio-demographics and inactive voter rates—whether localities pay attention to data quality such as missingness vary widely. These studies provide the foundation for the analysis we report in this paper.

There should be more research on recognizing and addressing problems stemming from the aggregation of decentralized sub-state databases. Decentralized election administration efforts, while attempting to enhance integrity, can affect political representation, for example, by determining who gets access to the ballot (Huber et al. 2021). Given that localities have a great deal of control over management practices that can affect representation, there is surprisingly little oversight about whether these practices are consistent, even within the same state. Research has argued that relatively simple means of oversight, like mandatory management training or list maintenance meetings, may not resolve differences in list maintenance and quality across local jurisdictions (Merivaki 2020).

Our approach adds to this important emerging literature on the quality of critical administrative data by systematically evaluating local patterns within the state and detecting anomalies. We start with data similar to that use by Kim, Schneider, and Alvarez (2019), in that we use repeated instances (or "snapshots") of a jurisdiction's voter registration database and implement repeated record linkage using state-provided voter IDs. We use the matched results to build multiple time-series that will help us assess database quality, such as the time-series of the number of new records, the number of records dropped, and the number of records that changed in key fields (e.g., address or partisan affiliation) between the snapshots. Unlike Kim, Schneider, and Alvarez (2019), though, we develop a more sophisticated statistical model to detect anomalies (on substantially larger statewide datasets) to ascertain whether a particular rate of change in any of these time series is a statistical outlier—and thus qualify for further investigation.

It is an important innovation to take the problem of anomaly detection in voter data to the state level. First, instead of focusing on a single county as in Kim, Schneider, and Alvarez (2019), the model looks for anomalies *across all of the counties in a state*. This gives us the ability to not only look for outliers over instances of the dataset (*over time*), but it also provides us with 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. This ability to search 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 so that their changes would be correlated. This allows our model to incorporate information about changes throughout the administrative data, improving the model's ability to detect abnormalities.

As an example to demonstrate the utility of our method, we apply it to public-release voter registration data from the 2020 election cycle in California. California

is a large and diverse state with 58 counties. California's voter registration database is gigantic, with approximately 22 million registered voters in 2020. Maintenance of this large dataset is complex, as there are many different state and local agencies with direct access to the data system. At the state level, the Secretary of State's VoteCal group, the state's Division of Motor Vehicles, and other state agencies can provide voter registration information to the statewide system. But each of the state's counties has access to the voter registration data system as well, conducting routine voter registration and file maintenance activities.

Unfortunately, there is little in the way of a deductive theory that can be used in this type of analysis. The forensic study of voter registration databases is relatively new. There are few previous studies that we can rely upon to help us draft hypotheses about what we might *expect* to see in the application of these methodologies to any particular state's voter registration data. This lack of deductive theory is an important rationale for the way we approach this problem. If we knew in advance where to look for anomalies, based on deductive theory, there would not be much reason for the development and application of anomaly detection methodologies like we use in this paper. We return to this discussion below.

3 Data and Methodology

We obtained the public-release voter registration and voting history data directly from the California Secretary of State. These data were provided weekly on DVD.[3] For this paper, given the interest in potential voter registration database error in the 2020 presidential election, we started to acquire data in early 2020—the data spans May 7th, 2020 to Nov 25th, 2020, the critical final months in the presidential election cycle when the likelihood of malicious registration might be great (and when unintentional administrative error might be problematic).[4] California's voter data management system, VoteCal, is a centralized voter registration database. The state describes VoteCal as "a single place for list maintenance functions" for the county elections officials. It is connected to state and county information systems.[5]

3 For information on how to make a Public Voter Registration Data Request (PVRDR) from the California Secretary of State's Office, see https://elections.cdn.sos.ca.gov/ccrov/pdf/2018/may/18100cik.pdf.

4 The snapshots were generated by the California Secretary of State's VoteCal office on May 7th, Jun 4th, Jun 8th, Jun 15th, Jun 22nd, Jun 29th, Jul 7th, Jul 13th, Jul 20th, Jul 28th, Aug 4th, Aug 10th, Aug 17th, Aug 24th, Aug 31st, Sep 9th, Sep 16th, Sep 23rd, Oct 1st, Oct 14th, Oct 21st, Oct 28th, Nov 12th, Nov 18th, and Nov 25th of 2020.

5 These include the County Election Management Systems (EMS), California Department of Corrections and Rehabilitation (CDCR), California Department of Public Health (CDPH), California Employment Development Department (EDD), and California Department of Motor Vehicles (DMV).

3.1 Matching Records in Snapshots of Voter Registration Database

We match the consecutive snapshots of voter registration data using unique voter IDs (RegistrantID) provided by VoteCal. We then identify voter records that were either added, dropped, or changed in the period between the extractions of two consecutive snapshots:

- **Added Records**: Records from snapshot t with voter IDs that cannot be found in snapshot $t-1$.
- **Dropped Records**: Records from snapshot $t-1$ with voter IDs that cannot be found in snapshot t.
- **Changed Records**: Records from snapshot t that have voter IDs been found in existing records in snapshot $t-1$, but at least one of the fields (e.g., Address, LastName) are changed during the period $t-1$ to t.

Formally, we express these change rates as follows, where $c \in \{58 \text{ counties in California}\}$, $t \in \{24 \text{ extraction dates between Jun 4th and Nov 25th}\}$, i is record indicator, and $k \in \{\text{Last Name, Residential Address, Date of Birth, VBM Status, Party}\}$:

$$\widehat{\text{Added}}_{c.t} = \frac{\text{Number of Added Records in County } c \text{ at Time } t}{\text{Average Total Records in County } c \text{ at Time } t-1 \text{ and } t} = \frac{\sum_c I_{\left(x_{i,t} \text{is added}\right)}}{(N_{c,t-1}+N_{c,t})/2}$$

$$\widehat{\text{Dropped}}_{c,t} = \frac{\text{Number of Dropped Records in County } c \text{ at Time } t-1}{\text{Average Total Records in County } c \text{ at Time } t-1 \text{ and } t} = \frac{\sum_c I_{\left(x_{i,t-1} \text{is dropped}\right)}}{(N_{c,t-1}+N_{c,t})/2}$$

$$\widehat{\text{Changed}}^k_{c,t} = \frac{\text{Number of Changed Records in County } c \text{ at Time } t \text{ in Field } k}{\text{Average Total Records in County } c \text{ at Time } t-1 \text{ and } t} = \frac{\sum_c I_{\left(x^k_{i,t} \text{is changed}\right)}}{(N_{c,t-1}+N_{c,t})/2} \quad (1)$$

3.2 Finding Anomalies in Voter Registration Changes

Having linked the voter records and obtained the rate of change quantities by 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, and the particular counties, where the rates of change are sufficiently large compared to other local entities at the same level as deemed an anomaly deserving further examination.

We want to use an approach for statistical anomaly detection that identifies counties and 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 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 California counties with potentially anomalous rates of change in the voter registration data, 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 California counties. Second, we provide a Bayesian multivariate analysis to model the change rates and check for counties with 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 following subsection, we provide the details for our Bayesian anomaly detection methodology.

3.3 A Bayesian Approach

For a principled detection of anomalies, we use a Bayesian model. It first estimates a model that best explains the variance within changes to the data (e.g., added rates, dropped rates, and changed rates), creates predictions from the posterior parameter distributions, and identifies deviations with 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*, whose changes over time are expected to be correlated. This use of a multivariate model to detect anomalies in voter list maintenance is one of the primary contributions of this paper.

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. Bayesian methods can combine prior information and data to derive posterior distributions of the parameters. In this study, we use the R package `brms` (Bürkner 2017) which is based on the `Stan` programming language (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 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 group 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**: Generate predictions for each dependent variable by making one random draw from the posterior distribution for each HMC iteration after the warm-up period,[6] and then computing the mean of the random draws.
- **Step 4**: Quantify the differences between the dependent variables and the means of model predictions by using *t*-scores from the hypothesis tests with the null hypothesis that *difference is zero*.
- **Step 5**: Identify the significant deviations that have *p*-values less than 0.05.

The multivariate model can be written as follows:

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

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

6 The number of random draws λ = *number of iterations after warm-up number of chains*. We recommend a $\lambda > 1000$ for efficient *t* tests. In our California application, we make $\lambda = 6000 \times 4 = 24{,}000$ random draws.

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

Each row of $Y_{n\times p}$ stands for a county × time combination, so that $n = n_{\text{county}} \times n_{\text{time}} = 58 \times 24 = 1{,}392$.[7] Columns of $Y_{n\times p}$, i.e., $[\boldsymbol{y}_1\ \boldsymbol{y}_2 \cdots \boldsymbol{y}_p]$ are $p = 7$ types of 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, $[\mathbf{x}_1\ \mathbf{x}_2]$ are group level variables {*County*, *Time*}. For the rest of the model, $\beta_{(r+1)\times p}$ are coefficients and $\varepsilon_{n\times p}$ is error structure:

$$\boldsymbol{\epsilon}_{n\times p} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \cdots & \epsilon_{1,p} \\ \epsilon_{2,1} & \epsilon_{2,2} & \cdots & \epsilon_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \cdots & \epsilon_{n,p} \end{bmatrix} = \begin{bmatrix} \epsilon_1' \\ \epsilon_2' \\ \vdots \\ \epsilon_n' \end{bmatrix}$$

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

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

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:

7 The data was received for 25 different dates, which translates into 24 different time periods where the data potentially changed.

$$\mathrm{ZIB}\left(y_{i,j};\alpha_{i,j},\mu_{i,j},\phi_{i,j}\right)=\left\{\begin{array}{ll}\alpha_{i,j} & \text{if } y_{i,j}=0\\ (1-\alpha_{i,j})B\left(y_{i,j};\mu_{i,j},\phi_{i,j}\right) & \text{if } 0<y_{i,j}<1\end{array}\right\}$$

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

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

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

$$\left\{\begin{array}{ll}E\left(y_{i,j}\right) & =\mu_{i,j}\\ \mathrm{Var}\left(y_{i,j}\right) & =\dfrac{\mu_{i,j}(1-\mu_{i,j})}{\phi_{i,j}+1}\end{array}\right.$$

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

$$\ln\frac{\Pr\left(\mu_{i,j}=1|x_i\right)}{1-\Pr\left(\mu_{i,j}=1|x_i\right)}=x_i\beta_j,$$

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

Once we have run our Hamiltonian Monte Carlo estimation (Duane et al. 1987; Neal 2011), 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 mean of model prediction ($\overline{\widehat{y}}_{i,j}$) can be found using the following hypothesis test:

$$\begin{array}{l}H_0: y_{i,j}=\overline{\widehat{y}}_{i,j}\\ H_a: y_{i,j}\neq\overline{\widehat{y}}_{i,j}\end{array}$$

4 Results

4.1 An Application to California Registration Data

In this section, we apply our multivariate Bayesian model to the California state-level voter registration data and discuss the potential anomalies. Readers interested in a visual introduction to the rates of change can see those in Appendix A in Supplementary material. The t scores of the hypothesis tests are shown in Appendix B in Supplementary material. t score higher than 1.96 indicates $y_{c,t}$ is significantly ($\alpha < 0.05$) different from model prediction $\overline{\widehat{y}}_{i,j}$ To keep the figures

concise and easy to read, we only plot outliers that are outside 3 × IQR or have t scores higher than 2.58 ($\alpha < 0.01$).

Our study on the California registration data finds that, out of 9,744 events,[8] 288 (3%) of the change events have been flagged as significantly ($\alpha < 0.05$) different from model predictions. This is not very surprising to those familiar with the heterogeneity in large, administrative datasets, especially voter files where many entities actively change the data via a hybrid system.

To aggregate these deviations into a more succinct picture, Figure 1 shows, out of 168 events (7 metrics × 24 snapshots), how many times a county deviated from the model prediction. There were 54 counties with at least one deviation and 23 counties with at least five deviations. Lake, Alpine, Stanislaus, San Diego, Mono had 10–13 deviations, while Modoc and Yolo had 15. Kern, which was the one with the highest number of deviations, exhibits change rates that are significantly ($\alpha < 0.05$) different from model predictions in 19 events. If our interest lies in seeing

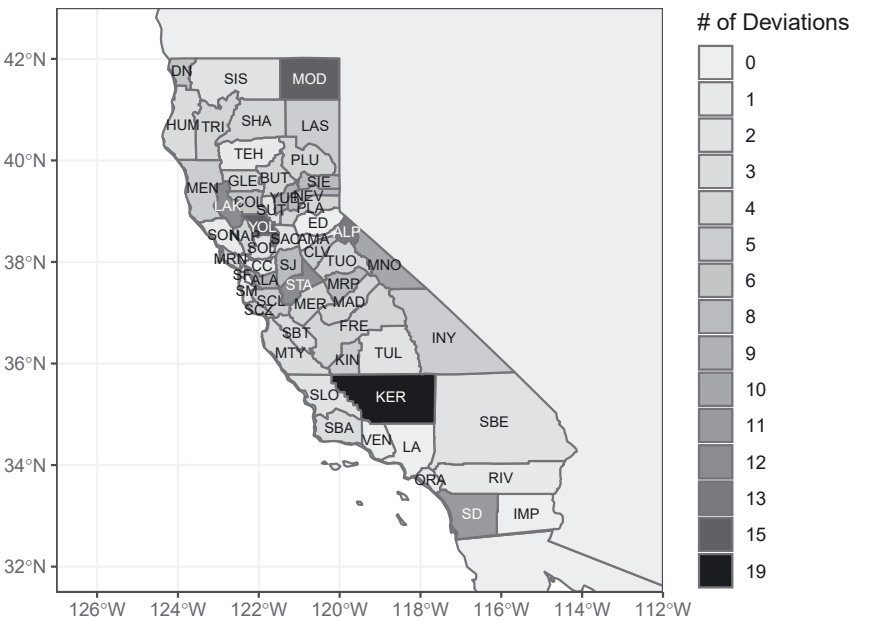

Figure 1: Number of deviations from model prediction in each county.

8 This is calculated as $n_{\text{county}} \times n_{\text{time}} \times n_{\text{variables}} = 58 \times 24 \times 7 = 9744$.

which counties have very different voter list maintenance results, we can sort the counties by this order, and Figure 1 can serve as a benchmark.

As we have warned in the introduction, this does *not*, per se, mean that there have been external or internal mishaps from incompetence or bad faith. It simply means that a state-level official may want to investigate the details of why some counties—or which particular dates—generate more outliers. It presents administrators with an opportunity to find out what causes high levels of heterogeneity, and whether the state should and/or can help resolve any problems.

To show, however, that the model performs better than a simple setup of an outlier cutoff, we show how the model can be validated using simulations in the next subsection.

4.2 Validating the Bayesian Method Using Synthetic Changes

In this section, we randomly modify the change rates to mimic errors and intrusions in the CA voter database, and provide a spectrum of "potential outliers" that span from mild to severe situations. This synthetic dataset allows us to investigate the performance of the Bayesian method along with the IQR method, and to reveal the fundamental difference between the two methods.

We will show that the Bayesian method can identify most of the "potential outliers" detected by the IQR method. In addition, for the synthetic changes that are ambiguous for both methods to find, the Bayesian method takes the nine-dimensional information (i.e., seven change rates in the model, and date and county-level group effects) into account and detects the synthetic changes that systematically deviate from the model prediction, while the IQR method can only focus on one of the nine dimensions and find the isolated changes that seem to be abnormal in only one of the metrics.

We use the following steps to make the synthetic changes and to implement the anomaly detection methods. First, we randomly select 200 change rates with no replacement from the CA change rates that were not identified as deviations in the previous application. Then for each selected change rate, we randomly apply one of ten manipulations (adding 0.1, 0.3, 0.5, or multiplying by 1.25, 1.5, 2, 4, 8, 16, 32) to assure that the modified data evenly cover the mild and severe situations. This is beneficial because we can have many synthetic change rates that are close to the thresholds where the Bayesian and IQR methods make decisions (t-score = 1.96, 1.5 × IQR), and it allows us to compare the behaviors/performances of the methods on the ambiguous change rates. Next, we split the 200 synthetic change rates into 50 groups and generate 50 synthetic data sets (i.e., each data set has four synthetic changes). Introducing four synthetic changes at a time avoids

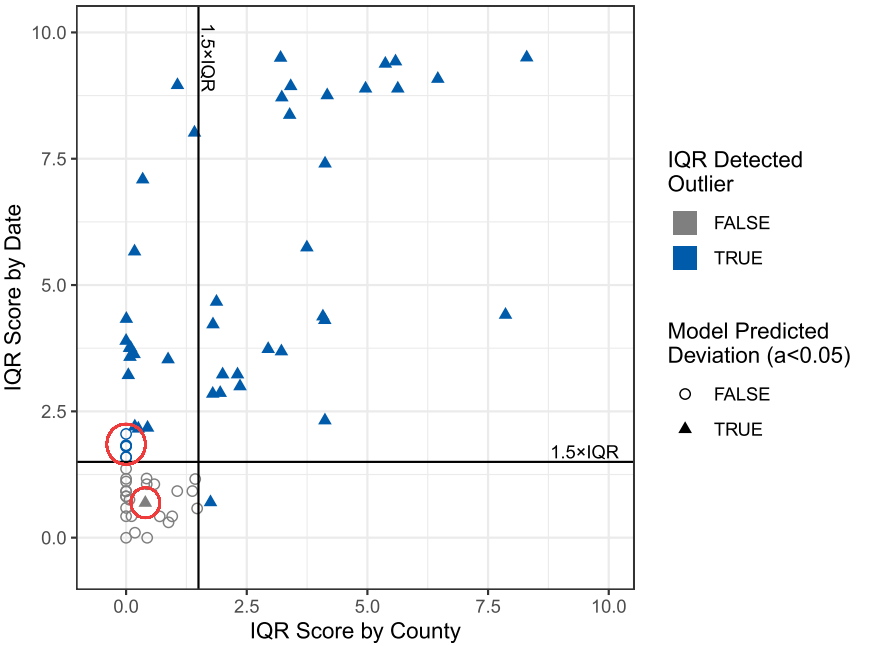

Figure 2: Deviations detected by the Bayesian method ($\alpha < 0.05$) and outliers detected by the IQR method ($1.5 \times$ IQR).

driving the average change rates up to the extent that some "outliers" might slip away from the anomaly detection methods. Lastly, we implement the Bayesian and IQR methods 50 times and plot the results in Figure 2.

As shown in Figure 2, to be able to visually compare the performances of the anomaly detection methods, we put the 200 synthetic changes on a two-dimensional plane with by-county IQR score as the x-axis and by-date IQR score as the y-axis. The IQR score $\Omega \geq 0$ is used to show how far away a change rate is from the first/third quantile in terms of interquartile range (IQR). Formally, it is defined as

$$\Omega^{\lambda}_{c,t,k} = \begin{cases} \dfrac{\text{quantile}\left(Y_{\lambda,k}, 0.25\right) - y_{c,t,k}}{\text{IQR}\left(Y_{\lambda,k}\right)} & y_{c,t,k} < \text{quantile}\left(Y_{\lambda,k}, 0.25\right) \\ \dfrac{y_{c,t,k} - \text{quantile}\left(Y_{\lambda,k}, 0.75\right)}{\text{IQR}\left(Y_{\lambda,k}\right)} & y_{c,t,k} > \text{quantile}\left(Y_{\lambda,k}, 0.75\right) \\ 0 & \text{other} \end{cases}$$

where $y_{c,t,k}$ is change rate of field k in county c at date t. $Y_{\lambda,k}$ is vector of change rates of field k in county c or at date t (i.e., $\lambda \in (c, t)$). For instance, a by-date IQR score

$\Omega^t_{c,t,k} = 3.1$ means the change rate $y_{c,t,k}$ is 3.1 × IQR away from the first or the third quantile of the change rate vector $Y_{t,k}$, whichever is nearer. It also means a $\phi \times$ IQR anomaly detection method with $\phi \geq 3.1$ would detect this change rate as an outlier.

The synthetic change rates are presented as points on this two-dimensional plane, with their by-county and by-date IQR scores as the coordinates. Although these synthetic changes were made to mimic errors and intrusions in the voter database, not all of them can be detected as outliers. Visually, if we put two lines ($y = 1.5$ and $x = 1.5$) in the figure, any change rates that are above the horizontal line ($y = 1.5$) or on the right side of the vertical line ($x = 1.5$) would be detected by the 1.5 × IQR method as outliers (blue points), and the rest of the points in the bottom-left quadrant would not be detected (grey points). This is reasonable as the grey synthetic change rates are so similar to the normal change rates and they are not abnormal enough to be identified from the natural variation. While it is inevitable that some errors and intrusions could be overlooked by the anomaly detection methods, their impact is limited.

In addition to marking the IQR outliers (1.5 × IQR) in blue, we show the deviations from the Bayesian model predictions as triangles. With the colors and shapes, we can easily compare the performance of the two anomaly detection methods. Assuming that the Bayesian method ($\alpha < 0.05$) works exactly the same as the IQR method (1.5 × IQR), we would see only grey dots in the bottom-left quadrant and blue triangles in the other parts of the figure. However, as highlighted by the red circles, there are inconsistencies between the two methods. In the upper red circle, six synthetic change rates are detected by the 1.5 × IQR method as their by-date IQR scores are larger than 1.5 (with a mean around 1.8), but not identified by the Bayesian method. The reason is although the six points are different from the other change rates of the same date (by-date IQR scores ≈ 1.8), they are similar to the change rates of the same county (by-county IQR scores ≈ 0). After integrating the nine-dimensional data (the county and time group effects plus the seven change rates), the Bayesian method re-weights the isolated mild anomalous (by-date IQR scores ≈ 1.8) and determines that the six change rates are not significant ($\alpha < 0.05$) deviations. In the lower red circle, one synthetic change rate ($\Omega^c = 0.4$, $\Omega^t = 0.69$) is detected by the Bayesian method ($\alpha < 0.05$) but not the IQR method (1.5 × IQR). This is because although the individual IQR scores are less than 1.5, the accumulative anomaly across multiple dimensions exceeds the threshold ($\alpha < 0.05$) of the Bayesian method. We argue that our Bayesian method fits better in election database anomaly detection as it detects mild systematical deviations (the triangle in the third quadrant) while reduces the false positives (the six circles in the first quadrant).

Figure 2 shows that the Bayesian method ($\alpha < 0.05$) can detect most of the "potential outliers" that are detected by the 1.5 × IQR method even though the two

methods are fundamentally different in handling multi-dimensional information. For the dataset in which fields are hypothetically correlated, such as the change rates in the CA voter database, the Bayesian method is preferred as it takes multi-dimensional information into account and makes aggregative judgments.

4.3 Probing the Reasons for Anomalies

Given that the model performs well, and we can find which values deviate significantly from model predictions, the natural step is now probing why these anomalies occur. In this subsection, we briefly delve into the county-level aggregation of the number of anomalies detected.

The literature suggests that we should consider the resources available for effective election administration, in particular the availability of resources from each county, and how much each jurisdiction spends on elections per voter (Hill 2012; Kimball and Baybeck 2013; Kropf et al. 2020; Montjoy 2010). We hypothesize that counties that can or are providing higher levels of resources to election administration may have lower rates of registration data anomalies. Better-resourced counties are more likely to have staff dedicated to working with their registration data, and to be routinely engaged in file maintenance. Another consideration is whether the county is participating in California's *Voter's Choice Act*, which allows counties to use universal voting-by-mail. It is possible that universal voting-by-mail counties might take additional steps to develop more accurate voter registration databases, and thus have fewer anomalies. It might also be the case that because VCA counties are making more extensive use of mailing services that they are getting data from their mail balloting efforts that helps them better maintain their registration data. Thus we hypothesize that VCA counties may have fewer anomalies in their registration data than non-VCA counties.

Table 1 shows two regressions, one in the simple linear form and another in the Poisson regression form to account for the count nature of the data, to see whether any administrative factors have caused higher frequencies of anomalies. The following variables are used: average election cost between 2017 and 2019 per voter, defined as the amount spent divided by the number of registered voters, county's revenue (1 million USD unit), and whether the county is participating in the *Voters Choice Act*. These are potential factors that we hypothesize may affect election administration performance.[9]

9 Two counties were left out due to lack of data: San Francisco and Fresno.

Table 1: County-level regression results to check for systematic anomaly generation.

	Dependent variable	
	Number of deviations from model prediction	
	OLS (1)	***Poisson*** (2)
Election cost per voter	0.134 (0.079)	0.020** (0.079)
Countywide revenue	−0.0001 (0.0002)	−0.0001 (0.00003)
VCA county	−1.990 (1.350)	−0.438* (0.171)
Intercept	3.310* (1.550)	1.390*** (0.164)
Observations	56	56
R^2	0.120	
Adjusted R^2	0.069	
Log likelihood		−167.000
Akaike Inf. Crit.		342.000
Residual Std. Error	4.100 ($df = 52$)	
F statistic	2.360 ($df = 3; 52$)	

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Given that there are only 56 observations in this dataset, we need to temper expectations regarding statistical significance. Figure 3 in particular shows first the distribution of election costs and then the scatterplot between election cost per voter and the number of outliers, with the linear regression line. Although the Breusch–Pagan test does not give a significant result ($p = 0.4$), note that we only have 56 observations here.

In both regressions, the election cost per voter is *positively* (if weakly) correlated with the number of deviations (linear regression $p = 0.095$ and Poisson

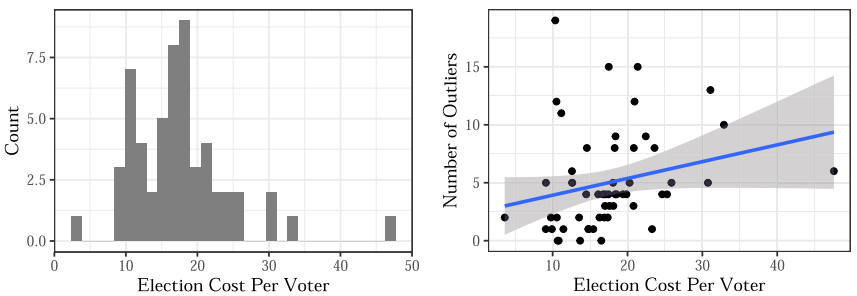

Figure 3: Distribution of election costs and the bivariate relationship between election costs and outliers.

regression $p = 0.006$). This hints that it is highly likely that there is a latent factor that is both driving the election cost and the number of anomalies upwards. What that latent factor is unknown, given our small dataset. However, it is another proof that investigating the cause of a high number of deviations is a worthy operation for the state as an upper-level control tower.

Again, it may be the case that the Bayesian model may need to be more permissive or need to be more tailored to California-specific practices. The current model is a proof of concept that can be tailored with more domain expertise and input from state-level and county-level election officials.

5 Conclusions

This paper argues that monitoring statewide voter databases is an under-emphasized and under-utilized tool in discussions of election reforms. Because the state voter data is comprised of sub-state data with varying list maintenance practices, it is a good practice to check whether changes at the county- or municipality-level are falling in step at the state level.

We developed a multilevel multivariate Bayesian model, and applied it to California's statewide voter data in 2020 to check whether counties generate similar changes across several key variables—such as the proportion of voters with changes to partisan affiliation—and observation periods. Our method and application adds to the growing literature on election forensics (Alvarez et al. 2020). However, previous studies of election forensics have focused largely on election outcome statistics (like turnout or candidate vote shares), and have not generally used flexible models (like our Bayesian approach) to incorporate a broad array of information to reduce the number of falsely identified anomalies. A good example of a related study is the methodology developed in Rozenas (2017), who uses a resampled kernal density approach to reduce the number of falsely identified anomalies in candidate vote share data. Our research provides direction for scholars and applied researchers who are concerned about reducing false positives in election forensic analyses.

We find that our model performs better than using a brute-force IQR method in detecting synthetically generated changes to real-world data. In addition, when applied to the actual data and aggregated at the county level, it can pinpoint certain counties that might be further studied. In fact, we find that the number of anomalies seems to be positively correlated with the average cost of elections per registered voter, hinting that there may be systematic inefficiencies that cause both a high cost and a greater number of anomalies.

Our approach to monitoring statewide voter registration data is of particular importance for the 17 U.S. states and territories that have "bottom-up" or "hybrid" voter registration systems, where local election officials play a role in list maintenance.[10] In all of these states and territories, the need to monitor for anomalous local variation in file maintenance is critical. Thus, we encourage election officials and researchers in those states to use our methodology to help monitor the integrity of registration data in those jurisdictions.

A metaphor that would be useful here is that of marathon organizers. The organizers are aware of the typical statistics that accompany a race—for example, the typical number of minutes that participants take to complete a 5k-race. While it is not wrong for some participants to lead or lag compared to the rest of the group, the organizers are constantly monitoring to see whether the ones lagging behind need medical assistance or the ones outpacing are deviating from the set course. A laissez-faire approach will make it difficult to ensure the goal of the race—the success and security of all participants.

Similarly, state-level election officials can start building data that can be added to statistical models like ours, and continuously monitor the model outputs. In some cases, deeper dives into why deviations from model predictions occur will be helpful. Such efforts will help protect the security and integrity of statewide voter databases, rather than assuming that all sub-state list maintenance efforts are by themselves sufficient. Efforts like these can also help insure that differences in voter list maintenance practices across counties do not make it more difficult for eligible citizens to obtain and cast their ballots.

References

117th Congress (2021–2022). 2021. *For the People Act of 2021*.

Alvarez, R. M. 2005. *Potential Threats to Statewide Voter Registration Systems*. Caltech/MIT Voting Technology Project Working Paper 40. Caltech/MIT Voting Technology Project.

Alvarez, R. M., J. Jonas, W. E. Winkler, and R. N. Wright. 2009. "Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project." In *Proceedings of the 2009 Electronic Voting Technology Workshop-Workshop on Trustworthy Elections*.

Alvarez, R. M., N. Adams-Cohen, S.-Y. Silvia Kim, and Y. Li. 2020. *Securing American Elections: How Data-Driven Election Monitoring Can Improve Our Democracy*. New York, NY: Cambridge University Press.

10 See the Policy Survey Table 1, pages 85–86 (U.S. Election Assistance Commission 2021). The states with bottom-up systems like California's are Arkansas, Connecticut, Illinois, Mississippi, Nevada, New York, Ohio, Tennessee, and Utah. Puerto Rico also has a bottom-up registration system.

Ansolabehere, S., and E. Hersh. 2010. *The Quality of Voter Registration Records: A State-By-State Analysis*. Report 6. Caltech/MIT Voting Technology Project.

Ansolabehere, S., and E. Hersh. 2014. "Voter Registration: the Process and Quality of Lists." In *The Measure of American Elections*, edited by B. C. Burden, and C. S. III, 61–90. New York: Cambridge University Press.

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

Bowler, S., T. Brunell, D. Todd, and G. Paul. 2015. "Election Administration and Perceptions of Fair Elections." *Electoral Studies* 38: 1–9.

Burden, B. C., D. T. Canon, K. R. Mayer, and D. P. Moynihan. 2012. "The Effect of Administrative Burden on Bureaucratic Perception of Policies: Evidence from Election Administration." *Public Administration Review* 72 (5): 741–51.

Bürkner, P.-C. 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*. Report 1. Caltech/MIT Voting Technology Project.

Chandola, V., A. Banerjee, and V. Kumar. 2009. "Anomaly Detection: A Survey." *ACM Computing Surveys* 41 (3): 15.

Chib, S., and E. Greenberg. 1995. "Understanding the Metropolis-Hastings Algorithm." *The American Statistician* 49 (4): 327–35.

Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Belin Heidelberg: Springer-Verlag.

Christen, P., and R. W. Gayler. 2013. "Adaptive Temporary Entity Resolution on Dynamic Databases." In *Advances in Knowledge Discovery and Data Mining, PAKDD 2013*, Vol. 7819, edited by J. Pei, V. Tseng, L. Cao, H. Motoda, and G. Xu, Lecture Notes in Computer Science, 558–69. Springer, Berlin, Heidelberg.

Christen, P. 2014. *Preparation of a Real Voter Data Set for Record Linkage and Duplicate Detection Research*. Working paper. http://users.cecs.anu.edu.au/~Peter.Christen/publications/ncvoter-report-29june2014.pdf.

Damlen, P., J. Wakefield, and S. 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): 331–44.

Duane, S., A. D. Kennedy, B. J. Pendleton, and R. Duncan. 1987. "Hybrid Monte Carlo." *Physics Letters B* 195 (2): 216–22.

Ferrari, S., and F. Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions." *Journal of Applied Statistics* 31 (7): 799–815.

Goel, S., M. Meredith, M. Morse, D. Rothschild, and H. Shirani-Mehr. 2020. "One Person, One Vote: Estimating the Prevalence of Double Voting in U.S. Presidential Elections." *American Political Science Review* 114 (2): 456–69.

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

Hastings, W. K. 1970. "Monte Carlo Sampling Methods Using Markov Chains and their Applications." *Biometrika* 57 (1): 97–109.

Hill, S. A. 2012. "Election Administration Finance in California Counties." *The American Review of Public Administration* 423 (5): 606–28.

Hoffman, M. D., and A. Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15 (1): 1593–623.
Huber, G. A., M. Meredith, M. Morse, and K. Steele. 2021. "The Racial Burden of Voter List Maintenance Errors: Evidence from Wisconsin's Supplemental Movers Poll Books." *Science Advances* 7 (8): eabe4498.
Kim, S.-Y. S., S. Schneider, and R. M. Alvarez. 2019. "Evaluating the Quality of Changes in Voter Registration Databases." *American Politics Research* 48 (6): 670–6.
Kimball, D. C., and B. Baybeck. 2013. "Are All Jurisdictions Equal? Size Disparity in Election Administration." *Election Law Journal* 12: 130–45.
Kropf, M., E. V. Jo Pope, M. Jo Shepherd, and Z. Mohr. 2020. "Making Every Vote Count: The Important Role of Managerial Capacity in Achieving Better Election Administration Outcomes." *Public Administration Review* 80: 733–42.
Liebschutz, S. F., and D. J. Palazzolo. 2005. "HAVA and the States." *Publius: The Journal of Federalism* 35: 497–514.
Merivaki, T. 2019. "Access Denied? Investigating Voter Registration Rejections in Florida." *State Politics & Policy Quarterly* 19 (1): 53–82.
Merivaki, T. 2020. ""Our Voter Rolls are Cleaner Than Yours": Balancing Access and Integrity in Voter List Maintenance." *American Politics Research* 48: 560–70.
Montjoy, R. S. 2010. "The Changing Nature … And Costs … of Election Administration." *Public Administration Review* 70: 867–75.
Mueller, R. S. 2019. *Report on the Investigation into Russian Interference in the 2016 Presidential Election*. Washington: Report, US Dept. of Justice.
Neal, R. M. 2011. "MCMC Using Hamiltonian Dynamics." *Handbook of Markov Chain Monte Carlo* 2 (11): 2.
Ospina, R., and S. L. P. Ferrari. 2012. "A General Class of Zero-or-One Inflated Beta Regression Models." *Computational Statistics & Data Analysis* 56 (6): 1609–23.
Pettigrew, S., and C. Stewart III. 2017. *Moved Out, Moved On: Assessing the Effectiveness of Voter Registration List Maintenance*. MIT Political Science Department Research. Paper No. 2018-1.
Rouan, R. 2020. "Private Election Vendors Face Little Oversight in Ohio Despite Ballot, Voter Purge Errors." *The Columbus Dispatch*. https://www.dispatch.com/story/news/politics/elections/2020/10/18/most-election-vendors-face-little-oversight-ohio-as-presidential-race-trump-biden-approaches/5977869002/ (accessed April 12, 2021).
Rozenas, A. 2017. "Detecting Election Fraud from Irregularities in Vote-Share Distributions." *Political Analysis* 25 (1): 41–56.
Shino, E., M. D. Martinez, M. P. McDonald, and D. A. Smith. 2020. "Verifying Voter Registration Records." *American Politics Research* 48 (6): 677–81.
Stan Development Team. 2019. *RStan: The R Interface to Stan*. R package version 2.19.2.
U.S. Election Assistance Commission. 2021. *Election Administration and Voting Survey 2020 Compre- Hensive Report*. In *A Report From the U.S. Election Assistance Commission to the 117th Congress, 633 3rd Street NW*. Washington: Suite 200, 20001.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/spp-2021-0016).