Detecting Unbalanced Election Fraud Approaches From Undervoting Irregularities*

Lion Behrens**
University of Mannheim
behrens@uni-mannheim.de

June 17, 2022

Abstract: I argue that in the presence of concurrent electoral contests on election day, ballot box stuffing and vote stealing can be detected from undervoting irregularities that emerge if protagonists of fraud fail to interfere into multiple races to equal extents. Conceptually, I introduce the distinction between balanced and unbalanced fraud approaches in the presence of several simultaneous electoral contests. Methodologically, I develop a non-parametric simulation model to detect and quantify systematic interference that stems from unbalanced fraud approaches. I illustrate the method on both (i) empirical data from recent Ecuadorian elections that report large extents of undervoting (ii) and simulated data for which the degree of fraud is known. While the empirical patterns that are inherent to Ecuadorian voting returns are well explained by systematic manipulation, alternative mechanisms that yield similar empirical patterns are outlined. This paper highlights the relevance of contextual information for the practice of election forensics in general and improves our understanding of undervoting irregularities in particular.

Keywords: *Election forensics; Electoral integrity; Undervoting irregularities; Monte Carlo simulation, Ecuador.*

*Paper prepared for the 12th Annual Conference of the European Political Science Association in Prague, June 23-25, 2022. I thank Sofía Garcia Durrer and Gloria Gerhardt for excellent research assistance as well as Oke Bahnsen and Viktoriia Semenova for helpful comments on prior versions of this manuscript. Substantial parts of this work were developed during a research stay at the *Latin American Social Science Institute FLACSO* in Quito, Ecuador. I thank Prof. Santiago Basabe Serrano for generously hosting me and organizing a series of expert interviews that greatly aided the development of this project.

**Lion Behrens is a Ph.D. Candidate at the Graduate School of Economic and Social Sciences (*GESS*) and a Research Associate at the Chair of Quantitative Methods in the Social Sciences and the Collaborative Research Center 884 "The Political Economy of Reforms", University of Mannheim, Germany.

Since the 'electoral revolution' (Norris 2014) lead to a dramatic increase in elections during the second half of the twentieth century, direct elections at national scale have been almost unanimously adopted by countries across the globe. Doubts about electoral integrity are, however, by no means restricted to authoritarian regimes and new democracies. In a sample of 57 countries from the World Values Survey collected between 2017-2022, merely 15.8% of respondents asserted that votes are counted fairly and election officials are fair in their country. Substantial doubts are voiced even among developed democracies, as this share of respondents rises to merely 25.1% if only the twelve OECD member states that are part of the sample are considered.[1] Electoral events worldwide are regularly followed by intense scrutiny of their level of integrity.

The field of election forensics employs statistical methods to detect anomalies in voting returns that are indicative of systematic irregularities. Conventionally, this body of research is driven by the assumption that fine disaggregated election results which stem from fraud-free processes inherit a range of numerical characteristics that hold globally over different electoral systems but are violated under manual alteration of the data. For instance, existing research has focused on identifying anomalous patterns in the distribution of raw vote totals (Mebane 2008; Beber and Scacco 2012; Medzihorsky 2015), the share of polling stations that report exactly coarse integer percentages for turnout and the winner's vote share (Kobak, Shpilkin and Pshenichnikov 2016; Rozenas 2017), and systematic clusters, skewness and kurtosis within the bivariate distribution of turnout and support rates (Myagkov, Ordeshook and Shakin 2009; Klimek et al. 2012).

A distinct approach to statistical election fraud detection lies in exploiting specific features of an electoral system that are inherent to selected country cases. Analyzing Mexico's 2010 gubernational elections, Cantú (2014) exploits the fact that within each electoral precinct, eligible voters are assigned to polling stations according to their childhood surname. As voting behavior is uncorrelated with voters' last name initials, Cantú identifies systematic interference from unexpected differences in turnout levels and vote shares across contiguous polling stations.

This paper contributes to the literature on election forensics by exploiting the administration of simultaneous electoral events for the statistical detection of election fraud. The conduction of parallel events gives rise to the phenomenon of 'undervoting irregularities', which occur if *the same polling stations officially report diverging turnout levels across different electoral contests* and hence

[1] Based on the Values Survey Wave 7 (Haerpfer et al. 2022).

less (more) overall votes are observed for some races than for others. While at each individual polling station, the share of valid, invalid and spoiled ballots might differ across electoral contests, the total number of turned out voters necessarily needs to be identical across events. Given that no electoral laws are in place that formally restrict access to some electoral contests[2], these discrepancies are either the result of administrative errors or a consequence of fraudulent interference.

I argue that in the presence of concurrent electoral contests on election day, ballot box stuffing and vote stealing can be detected from undervoting irregularities that emerge if protagonists of fraud fail to interfere into multiple races to equal extents. Conceptually, I introduce the distinction between balanced and unbalanced fraud approaches in the presence of several simultaneous electoral events. Methodologically, I develop a non-parametric method of fraud detection building on the fact that if undervoting irregularities stem from administrative or human errors, discrepancies in turnout levels are unrelated to the winner's vote share. I illustrate the method on both (i) empirical data from recent Ecuadorian elections where undervoting irregularities are widespread (ii) and simulated data for which the degree of fraud is known.

This article makes two contributions to research on electoral fraud. First, I coin the conceptual distinction between balanced and unbalanced fraud approaches that enhances our understanding of how agents of fraud behave given that concurrent electoral events are taking place. Second, I contribute to the growing literature on statistical tools to detect fraudulent interference from numerical characteristics in fine-graded voting returns.

The remainder of this article is organized as follows. The next section explains the phenomenon of undervoting irregularities and showcases them based on national- and local-level data from Ecuador. Section 3 introduces the conceptual distinction between balanced and unbalanced fraud approaches across multiple electoral events. The subsequent section outlines a non-parametric simulation model estimating the degree of unbalanced fraud that is present. Lastly, I apply this method to three Ecuadorian elections in 2017 and 2019 and simulated elections for which the degree of fraud is known. While the empirical patterns that are inherent to Ecuadorian voting returns are well explained by systematic manipulation, an alternative mechanism which does not evoke fraud is outlined that would result in a similar empirical picture.

[2] An example of restrictive electoral laws would be underage voters or non-citizen residents only being eligible to vote in local elections but not in national contests that are conducted side-by-side.

A Motivating Example

Electoral History in Ecuador

Many countries hold parallel electoral contests on voting day. In Latin America alone, 60 out of the last 95 elections at national scale were conducted alongside at least one parallel contest.[3] To motivate the idea behind the method, I examine election data from the country of Ecuador.

During most of the first part of the twentieth century, Ecuador's electoral history was deeply permeated by institutionalized manipulation administered by the Radical Liberal Party (PLRE, de la Torre 2015), whose rule included practices such as restricting voting rights of marginalized groups, intimidation of opposition supporters and the alteration of final vote counts on ballot day. After the liberal party's main competitor Velasco Ibarra's fifth non-consecutive presidency ended in 1972 with a military coup, the country experienced a—comparatively short— period of military rule with no national elections conducted before in 1979 power was handed over the constitutionally elected civilian Social Democrat Jaime Roldós Aguilera.

The 1978 electoral reform granted illiterates—which formed a large part of the country's indigenous population who *de facto* have been excluded from the right of suffrage— the right to vote. Ever since, elections in Ecuador did formally function as a legitimate process to select legislative representatives and public officials. After the 1979 handover of power, Ecuador's history of democratization was marked by a large series of presidential downfalls which were tightly coupled to a number of economic crises often triggered by fluctuations of world oil prices. Steady shifts between governments favoring liberal free-market economics and left-winged platforms fostering social equality and protectionist measures characterized Ecuadorian politics up until the beginning of the 21st century. These constant shifts lead to a total number of twelve different presidents in the period between 1979-2007, out of which only few could regularly end their presidential term. During these decades, the conduct of elections often fulfilled the mere purpose of officially restoring a delicate power balance that was continuously disrupted by repeated *coups d'état*.

The period of repeated party system collapse came to an end when in November 2006, Rafael Correa Delgado, an independent leftist with no partisan base, was elected president and went on

[3] Source: Own research by the author. Time frame covered: 2009-2020. Countries covered: Argentina, Belize, Bolivia, Brazil, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Venezuela, Chile, Urugay.

to rule the country over three consecutive terms until 2017. Correa, a close ally of Venezuelan socialist leader Hugo Chavez, rapidly gained popular support implementing a platform that concentrated broad powers in the hands of the president, restored national control over the country's foreign owned oil industries and used flourishing oil revenues to implement extensive social spending alongside free secondary and post-secondary education. The second half of Correa's presidency was characterized by an increasingly authoritarian style, which—backed up by a series of constitutional referendums that received mass public support— witnessed a range of political reforms undermining the independence of the judicial system, growing control over media content by the government and the persecution of political opponents. As political institutions became more and more aligned with the president's *Correismo*, the Ecuadorian discourse about election integrity, while continuously revolving around voting day inconsistencies, was augmented by a second dimension. Political observers increasingly criticized Correa's systematic politicization of electoral institutions such as the National Electoral Commission (CNE)—which was packed with Correa supporters—and his extensive presidential control over the country's main media outlets, which failed to provide a level playing field for political competition.

The discourse about election integrity escalated in the 2017 presidential election, when Correa-endorsed successor Lenín Moreno faced the Guayaquil-based liberal banker Guillermo Lasso in the runoff vote held on April 2. After several major exit polls during election night had predicted a win for Lasso who already declared victory and an end to 13 years of *Correismo*, the CNE declared Moreno to be the winner over a small margin of less than 2 percent of the votes. As a consequence, the country was shook by waves of protests that lasted several weeks, although several recounts, predominantly in the region of Guayas, reassured Moreno's victory.

Today, Ecuadorian elections are characterized by repeated large scale protests that aim to question the legitimacy of election results, a large number of absentees (1 out of every 5 eligible voters) although the electoral system implements compulsory voting, and a large number of purposely spoiled ballots among those that do turn out (up to 17.9 percent in the country's most recent 2021 election) demonstrating the population's large distrust in the electoral process.

Undervoting Irregularities

Ecuadorian elections are regularly conducted as general elections in which multiple electoral races for different types of offices (*dignidades*) are conducted side-by-side. For instance, in the 2017 elections held on February 19, voters directly elected (i) the head of government in a presidential race, (ii) the members of the country's national assembly, (iii) parliamentary members of 24 regional assemblies, (iv) Ecuador's five national representatives for the Andean parliament—the deliberative body of the Andean community—and (v) cast votes in a nation-wide referendum prohibiting politicians and civil servants to hold bank accounts in countries with preferential tax regimes and low tax jurisdictions. On election day, voters are assigned to different polling stations according to their registered address. At the voting localities, voters get handed out different ballots corresponding to different electoral contests, which are inserted into separate ballot boxes. At the counting stage, votes for different electoral contests get tabulated in separate vote tally sheets (*actas*). The Ecuadorian electoral process has been described as highly complex by international election observers monitoring democracy (EU 2009), who have voiced for a simplification of the procedures.

Ecuadorian electoral returns are marred by empirical inconsistencies which are sometimes referred to as 'undervoting irregularities'. This peculiarity refers to the phenomenon of individual polling stations reporting diverging numbers of turned out voters across different electoral contests that are conducted side-by-side on voting day. Naturally, it is expected that some voters cast valid votes for some contests and submit empty or spoiled ballots for other electoral races depending on factors such as their saliency, perceived competitiveness or regional and political scope. However, it holds that across all electoral races, the same polling stations must report the same number of voters that turned out in the first place.

The left and right panel of Figure 1 show an empirical example of an undervoting irregularity for the Ecuadorian local elections conducted on March 24, 2019.[4] Depicted are two vote tallies (*actas*) from the same voting table (*junta receptora*) for two different electoral contests at a polling station in the municipality of Sidcay in Cuenca, Azuay. The left tally shows results for the election of provincial representatives (*prefectos*). Out of 350 eligible voters who are registered for this polling

[4]Material acquired digitally from the National Electoral Commission (CNE) on January 20, 2020 in Quito, Ecuador.

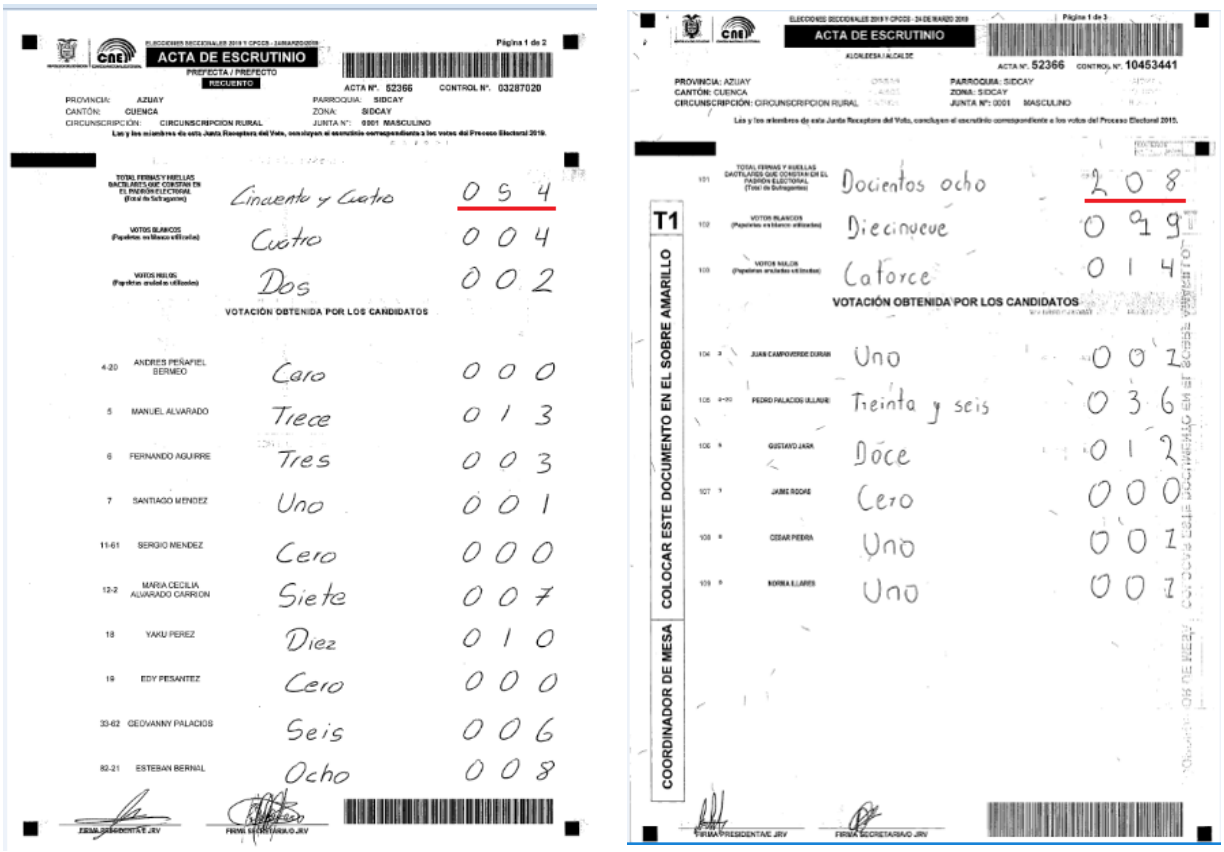


Figure 1: Irregularities on vote tallies. An empirical example of an undervoting irregularity in the Ecuadorian local elections on March 24, 2019 from the municipality Sidcay in Cuenca, Azuay. The left tally depicts the vote tabulation for provincial representatives (*prefectos*). The right tally tabulates votes for the city mayor (*alcalde*). The documented number of turned out voters is underlined in red and differs across both electoral races at the same polling station.

station, a total of 54 ballots have been observed including four blank and two null votes. The right tally shows results for the election of the city mayor (*alcalde*), for which a total of 208 votes have been counted including nineteen blank and fourteen null votes. This documents a substantial irregularity as the total number or received ballots for each contest must necessarily be the same on both tallies. The magnitude of the inconsistencies is a multiple of the vote distances between the individual candidates.

Figure 2 maps the extent of undervoting irregularities across more than 39,000 Ecuadorian polling stations for two of the recent elections which publicly have been called into question. Discrepancies in the documented number of turned out voters emerge in up to 2,980 (1,528) out of a total of $n = 39,319$ ($n = 15,857$) polling stations in 2017 (2019) at which parallel contests were

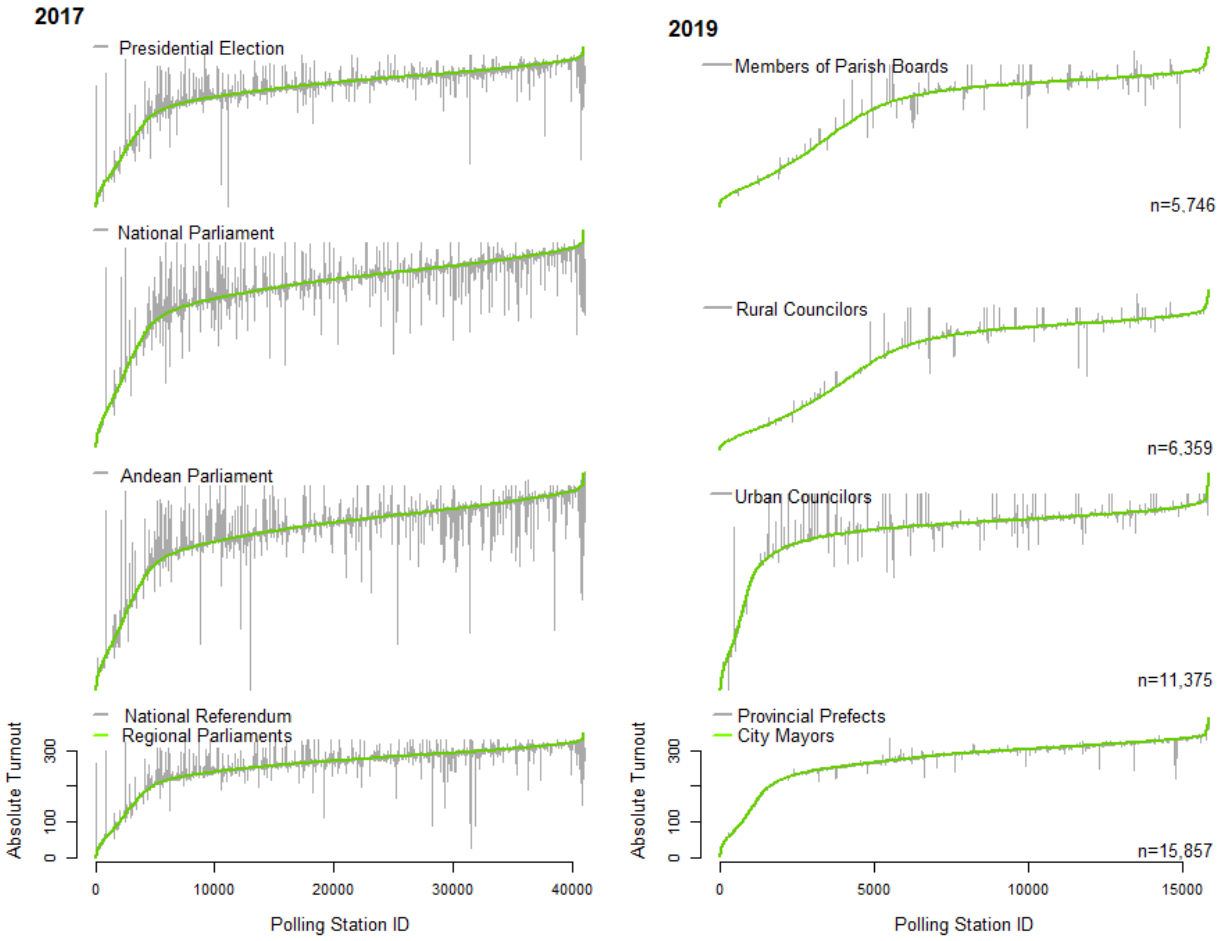


Figure 2: Undervoting irregularities across elections. Undervoting irregularities mapped across polling stations for five different electoral contests. Left panel: General Elections 2017. Right Panel: Local Elections 2019. Green lines indicate the absolute number of documented turned out voters for the election of state-level members of parliament (2017) and city mayors (2019). Grey lines indicate the absolute number of documented turned out voters in parallel elections at the same polling stations. Sample sizes differ for the 2019 elections as not all regional contests have been held in all provinces. Empirical patterns are insensitive to a change in baseline elections (green).

held. The magnitude of the discrepancies calls for the question whether these are due to administrative errors and the carelessness of low-level election officials or systematic manipulation on election day.

Election Fraud in the Presence of Concurrent Electoral Events

Balanced and Unbalanced Election Fraud

While multiple election day irregularities such as voter intimidation, electoral violence or the maladministration of polling stations have been observed in Ecuadorian elections (see for instance the Electoral Observer Mission reports OAS 2008, Carter Center 2007, EU 2009), this paper focuses particularly on ballot fraud. Ballot fraud is of central interest for large accounts of the literature on election integrity. Perhaps the most direct documentation of manual alterations of vote counts is delivered by Cantú (2019), who is providing a remarkable account of scanned vote tally-sheets from Mexico's 1988 presidential election which inherit crossed-out and altered numerals that are most prevalent in polling stations where the opposition was not present. Analyzing electoral returns from Russian presidential and Duma elections between 2000 and 2012, Kobak, Shpilkin and Pshenichnikov (2016) and Rozenas (2017) outline the specific mechanism of rounding fraud where vote shares for the favored candidate are simply rounded up at the vote tabulation stage, and both present statistical models that show that the number of exactly coarse integer vote-shares (e.g. 0.6, 0.65, 0.7) is significantly larger than expected by pure chance. Other methodological work on ballot fraud can be found in Beber and Scacco (2012), Myagkov, Ordeshook and Shakin (2009), Klimek et al. (2012) and Callen and Long (2015). There are two forms of ballot fraud that inflate (reduce) the number of turned out voters and thus are connected to undervoting irregularities: stuffing the ballot box with pre-prepared ballot papers or illegally withdrawing valid ballots from a race.

An explanation of the emergence of undervoting irregularities that can be put forward without the need of evoking systematic manipulation is that these might be manifestations of administrative challenges, human errors or the carelesness of electoral authorities during the voting or counting stage as polling places often fall short in personnel and material needed for faultless election conduct. Extensive election observation reports of the Organization of American States (OAS) and the European Union do a great job in describing the difficulties that Ecuadorian low-level election officials have faced, such as material arriving several hours after polling is supposed to start or shortage of electric light during the vote counting stage on election night (OAS 1998). Another explanation of undervoting irregularities is that some voters simply do not receive or hand back ballots for those races that they do not intend to vote for, either because election officials themselves

are misinformed about valid procedures, voters are misinstructed at the voting booth or due to their own carelessness. This reasoning is especially plausible when taking into account statements by Electoral Observation Missions stating that a substantial share of low-level election officials did not receive sufficient training to carry out their election-day duties (Carter Center 2007; EU 2002, 2009; IRI 2003), even citing individual cases where volunteers waiting in queue were spontaneously mobilized to help out on election day (OAS 2006, 2008).

In the case of Ecuador, the undervoting irregularities are indicative of a novel type of fraud that has not been described in the literature on election integrity. If low-level election officials or unauthorized individuals entering polling stations during voting or the counting stage selectively remove votes or add pre-prepared ballots to some of the ballot boxes, undervoting irregularities arise if ballot boxes concerning different electoral races are affected to unequal extents. An interview with a high-level representative of the Central Electoral Commission in Quito, Ecuador supports this view.

> *"What is happening? At certain tables, under the carelessness of the electoral authorities, they physically remove a number of votes, which constitutes an act of probabilistic fraud. If I know that 'Pedro' wins at Table A, I take one hundred votes from the ballot box. Most certainly, I will remove the majority of votes from 'Pedro'. I'll also remove votes from the rest of the candidates, but probably 'Pedro' is going to suffer the most from it. If I do this at ten tables, I achieve a relevant effect.*
>
> *It's simply groups of thieves [...] like the ones you see on the street. They tear the ballots out."*

This anecdote leads us to define two different strategies of fraud that can be present given that multiple electoral contests are conducted simultaneously: Balanced and unbalanced fraud. In a balanced fraud approach, all races are altered to exactly equal extents. That is, for every vote that is added to (or removed from) one electoral race, another vote is added to (removed from) all parallel contests, even those that are not of primary concern for the agents of fraud or their principals. In an unbalanced fraud approach, these numbers differ and undervoting irregularities emerge.

In the case of Ecuadorian elections, there are straightforward reasons substantiated in repeated reports of international Election Observation Missions to argue that undervoting irregularities are due to poor resources of administrative staff on election day combined with a complex voting system that conducts several contests side-by-side. Likewise, the sheer extent of turnout discrepancies and

anecdotes such as the one above speak in favor of a different mechanism in which undervoting irregularities stem from unbalanced fraud approaches in which ballots are selectively added or removed. This paper outlines a statistical approach to estimate the degree of undervoting irregularities that is due to unbalanced election fraud.

Undervoting Irregularities in the Absence of Unbalanced Fraud

In order to design a method that detects and quantifies unbalanced fraud approaches from undervoting irregularities and the winning candidate's (or party's) vote share, it is first important to understand that there is no expectation of a statistical relationship under human or administrative errors such as misinformed election officials, misinstructed voters, a loss of votes or miscounting. Before I describe a method for fraud detection, I hence first line out two properties of undervoting irregularities under random errors: In expectation, observed vote shares for all candidates are equal to their true vote shares even in the presence of excessive errors and there is no statistical association between the extent of undervoting irregularities and the winning candidate's vote share across polling stations.

Let N_i denote the number of eligible voters across $i = 1, ..., n$ polling stations. $\mathcal{T}_i \in [0, N_i]$ denotes the absolute number of turned out voters for a particular electoral race of interest and the share of votes the winning candidate (party) received is denoted by $p_i \in [0, 1]$. Across all polling stations, observed turnout levels $\mathcal{T}_i$ and winner's vote shares p_i can be decomposed as

$$\mathcal{T}_i = \mathcal{T}_i^* + \mathcal{T}_i^\epsilon \qquad \mathcal{T}_i^\epsilon \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

$$p_i = \frac{V_i}{T_i} = \frac{\mathcal{T}_i^*}{\mathcal{T}_i^* + \mathcal{T}_i^\epsilon} \underbrace{\frac{V_i^*}{\mathcal{T}_i^*}}_{p_i^*} + \frac{\mathcal{T}_i^\epsilon}{\mathcal{T}_i^* + \mathcal{T}_i^\epsilon} \underbrace{\frac{V_i^\epsilon}{\mathcal{T}_i^\epsilon}}_{\epsilon_i} \tag{2}$$

where $\mathcal{T}_i^*$ is the true number of turned out voters, $\mathcal{T}_i^\epsilon$ is the absolute number of discrepant votes from the true value either resulting from errors or fraud, V_i^* is the true absolute number of votes cast for the winner and V_i^ϵ is the number of votes for the winner among all lost or miscounted votes and V_i is the number of votes for the winner that is ultimately observed. Under the absence of fraud, the dispersion parameter σ is purely determined by structural factors such as the training of election officials and administrative or election day hurdles. Let us first consider the case in which

undervoting irregularities exclusively emerged simply because less votes were cast for a particular electoral contest as turned out voters did not receive or hand back all relevant ballots and there is no miscounting of those ballots that have been received. In this case, the error terms $\epsilon_i (i \in \{1, ..., n\})$ in (2) simply reduce to zero and observed vote shares are equal to true vote shares as $V_i^{\epsilon} = 0$.

A second scenario is given by ballots not getting accounted for, lost or miscounted although these were cast by turned out voters. Using the loss of votes as a working example, it is intuitive that $\epsilon_i = \frac{V_i^{\epsilon}}{\mathcal{T}_i^{\epsilon}}$ consists out of a subset of the true votes, and hence it is straightforward that the erroneous votes themselves can be written as a function of the true votes. Let us imagine that for each individual polling station i, there are $j (j \in \{1, ..., J\})$ different hypothetical scenarios in which true votes can be lost, and

$$\epsilon_i^j = \frac{V_i^{\epsilon,j}}{\mathcal{T}_i^{\epsilon,j}} = \frac{V_i^*}{\mathcal{T}_i^*} + \xi_i^j \qquad\qquad \xi_i^j \sim \mathcal{N}(\mu_i, \sigma_i^2). \tag{3}$$

ϵ_i^j is denoted as the share of votes for the winner among all lost (or miscounted votes) at a particular polling station, which varies across J hypothetical realizations. In clean elections, not accounted, miscounted or lost votes constitute a truly random sample of the true votes and hence $\mu_i = 0$, $E[\xi_i] = 0$ and $E[\epsilon_i] = \frac{V_i^*}{\mathcal{T}_i^*}$. This means that the expected portion of votes that were cast for the winner among all lost or miscounted votes is equal to the portion of votes cast for the winner among all votes that were originally cast. Reformulating (2), we can now straightforwardly derive that at each polling station, the miscount or loss of votes—in expectation—affects all candidates proportionally to their electoral strength in clean elections as

$$\begin{aligned} E[p_i] &= \frac{\mathcal{T}_i^*}{\mathcal{T}_i^* + \mathcal{T}_i^{\epsilon}} \frac{V_i^*}{\mathcal{T}_i^*} + \frac{\mathcal{T}_i^{\epsilon}}{\mathcal{T}_i^* + \mathcal{T}_i^{\epsilon}} E\big[\frac{V_i^{\epsilon}}{\mathcal{T}_i^{\epsilon}}\big] \\ &= \frac{\mathcal{T}_i^*}{\mathcal{T}_i^* + \mathcal{T}_i^{\epsilon}} \frac{V_i^*}{\mathcal{T}_i^*} + \frac{\mathcal{T}_i^{\epsilon}}{\mathcal{T}_i^* + \mathcal{T}_i^{\epsilon}} \frac{V_i^*}{\mathcal{T}_i^*} \\ &= \frac{V_i^*}{\mathcal{T}_i^*}. \end{aligned} \tag{4}$$

Equation (4) shows that the vote share that is expected to be observed at a particular locality is equal to the true vote share even in the case of excessive amounts of administrative or human errors.

Furthermore, let u_i be the extent of undervoting observed at one particular polling station when comparing the election of interest to a baseline electoral race and let $u_i = \frac{|\mathcal{T}_i^{\epsilon}|}{\mathcal{T}_i}$ be the share of votes

that are discrepant to a baseline race among the overall number of votes that have been observed in the main race. Across all polling stations, the covariance between the winner's vote share and the extent of undervoting is defined by

$$Cov(p,u) = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(u_i - \bar{u})}{n} = \frac{\sum \left(\frac{V_i}{T_i} - \frac{1}{n}\sum \frac{V_i}{T_i}\right) \times \left(\frac{|\mathcal{T}_i^{\epsilon}|}{\mathcal{T}_i^{*}+\mathcal{T}_i^{\epsilon}} - \frac{1}{n}\sum \frac{|\mathcal{T}_i^{\epsilon}|}{\mathcal{T}_i^{*}+\mathcal{T}_i^{\epsilon}}\right)}{n}. \tag{5}$$

From Equation (5), it follows that even if large amounts of undervoting irregularities are present that are due to human errors or electoral maladministration, the extent of undervoting is unrelated to the winning candidate's (party's) vote share in expectation as

$$\lim_{T_i^{\epsilon}\to\infty} Cov(p,u) = 0. \tag{6}$$

Figure 3 illustrates this point using artificial election results for concurrent events simulated under the protocol which I outline in the subsequent section. The properties of undervoting irregularities in the absence and presence of systematic manipulation can be nicely illustrated by visualizing the conditional distribution of winner's vote shares separately for different groups of polling stations with different levels of undervoting. In clean elections ($S = 0$), the distribution of winner's vote shares is unrelated to the extent of undervoting as vote shares homogeneously vary around their mean value independently of the extent of irregularities that is observed. When unbalanced fraud is incorporated, distributions are inflated in their upper tail as polling stations document higher levels of irregularities ($S = 0.2$ and $S = 0.4$) and take on substantially above-average values as the amount of fraud that is introduced becomes extreme ($S = 0.8$).

From Equation (6), it is tempting to calculate a linear model and infer fraud from a hypothesis test on the relationship between undervoting discrepancies and winner's vote shares. The main shortcoming of a linear model is that the produced measure of association is hard to interpret for the purpose of statistical fraud detection and thus would reduce inferences to a simple significance test. Therefore, I now outline a non-parametric simulation method that is designed to estimate a substantive quantity of interest—the share of polling stations with undervoting irregularities at which unbalanced fraud is conducted. The goal is to verify the validity of the quoted anecdote and to estimate the prevalence of unbalanced election fraud across Ecuadorian elections.

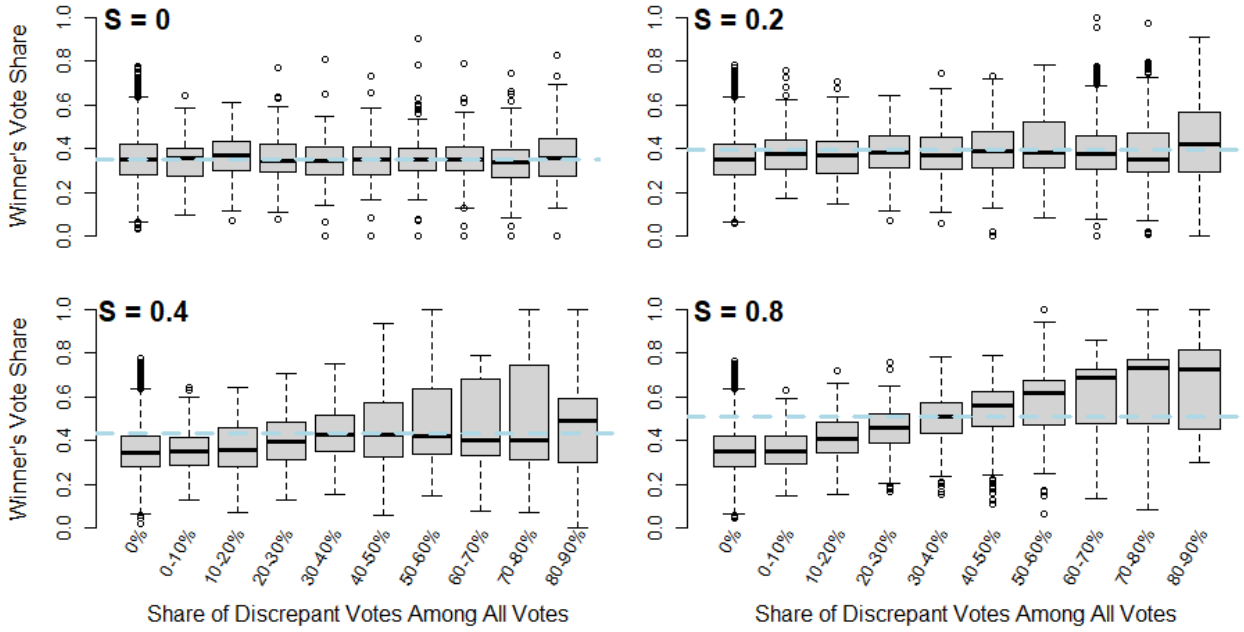


Figure 3: The relationship between undervoting irregularities and winner's vote shares (simulated). Boxplots summarize winner's vote shares in artificial data generated under the simulation model outlined in Section 4. Data was simulated for 40,000 polling stations out of which 3,000 were assigned undervoting irregularities. S describes the portion of polling stations with undervoting irregularities at which unbalanced fraud was executed. The share of discrepant votes among all votes is defined as $u_i = \frac{|\mathcal{T}_i^{\epsilon}|}{\mathcal{T}_i}$. The dashed blue line reports the average vote share of the winner (in favor of which results were altered) across all polling stations.

A Simulation Model to Detect Unbalanced Election Fraud

The goal of the following model is to estimate the share of polling stations with undervoting irregularities at which unbalanced fraud is conducted. In order to make statistical inference on the presence of systematic irregularities across simultaneous electoral events, we first need to specify a main electoral race for which this share is estimated and a baseline electoral event to which discrepancies in turnout are observed. As a general intuition, the non-parametric simulation model detects unbalanced election fraud by (i) simulating a range of artificial elections which mimic the main race that are either clean or manipulated to different degrees, (ii) quantifying the average distance between the empirical data and each set of simulated elections, (iii) and finding the set of artificial elections that–in expectation—minimizes the distance to the empirical data. The fraud parameter that was used to construct this set of artificial elections serves as the estimate of fraud.

Stochastic Process of Concurrent Elections

I model the two concurrent electoral events as

$$T_i^* \sim \text{Binomial}(N_i, t_i^*), \tag{7}$$

$$T_i^\epsilon = \begin{cases} 0 & \text{if } D_i = 0 \\ \text{Norm}(0, \sigma) & \text{if } D_i = 1 \end{cases} \tag{8}$$

$$V_i^* \sim \text{Binomial}(T_i^* + T_i^\epsilon, v_i^*), \tag{9}$$

for each polling station $i = 1, ..., n$. The absolute number of turnout T_i^* is set as the turnout in the baseline race observed at each locality and is defined as a binomial draw where the population size is the number of eligible voters at polling station i (the number of people in the voter register that have been attributed to a particular locality) with the polling station-specific success probability t_i^*. D is an indicator variable documenting whether undervoting is observed at a particular polling station. Turnout discrepancies T_i^ϵ to the main electoral race are set to 0 if $D_i = 0$ and defined as a draw from a normal distribution with a mean of zero and standard deviation σ if $D_i = 1$. This means that most turnout discrepancies take on small values, while the probability for larger discrepancies is decreasing. The absolute turnout for the main race then—by implication— is $T_i = T_i^* + T_i^\epsilon$. If T_i^ϵ takes on a positive value, this means votes are added to the main electoral race. If T_i^ϵ is negative, less overall votes are observed in the main race than in the baseline event. The number of people who vote for the overall winning candidate (party) V_i^* in the main race is a binomial draw from the population size $T_i = T_i^* + T_i^\epsilon$ (the number of observed votes at a particular polling station) with the success probability v_i^*.

To arrive at a fully specified stochastic process of two concurrent elections, what is missing is to parameterize the functions producing $\{t_i^*, v_i^*\}$ which represent the unknown distributions of polling station-level turnout and winner's support rates. Since these success probabilities necessarily fall in the $[0, 1]$ interval, it is intuitive to model these as beta distributions

$$t_i^* \sim \text{Beta}(\alpha^t, \beta^t), \tag{10}$$

$$v_i^* \sim \text{Beta}(\alpha^v, \beta^v), \tag{11}$$

where $\{\alpha, \beta\}$ are scale and shape parameters.

The stochastic process of elections that underlies my simulation model generally relies on two assumptions. Evidently, it assumes that the data generating functions which produce observed turnout T_i^* and winner's votes V_i^* are described reasonably well by binomial draws with success probabilities parameterized by beta distributions and that undervoting discrepancies T_i^ϵ are described by a normal distribution. This assumption is highly reasonable as data generated from these distributions resembles empirical data very closely and has been shown to hold equally well for other country contexts (see Rozenas 2017). Appendix Section A1 compares data simulated from this stochastic process to empirical data from Ecuador. A more subtle assumption is that empirical votes from the parallel contest that is set as the baseline race can be used to model the (latent) number of turned out voters and number of votes for the winner in the main electoral race T_i^*, V_i^* from Equations (1)-(2) before undervoting irregularities are introduced. It is important to note that this step does not assume that the baseline race itself is fraud-free. Rather, in the following, I present an approach to reverse-engineer the extent of *unbalanced* fraud between both races that is a consequence of intervening into both electoral races to unequal extents. This is not equal to the overall degree of fraud that might be present in the electoral data and does not assume that any of the used data was not manipulated at all. Rather, it exploits the fact that several simultaneous electoral events are taking place and infers the degree of unequal manipulation between the two.

A Non-Parametric Simulation Model

The non-parametric simulation model executes the following steps:

1. Set the overall number of polling stations n, the eligible voters per polling station N_i, and the number of polling stations with undervoting discrepancies $n^U = \sum D_i$ to their values in the empirical data.

2. Estimate $\alpha^t, \beta^t, \alpha^v, \beta^v, \sigma$ from the empirical data.

3. Sample values for T_i^*, T_i^ϵ from the distributions defined in (7)-(8). Sample values for V_i^* from Binomial(T_i^*, v_i^*) without incorporating undervoting yet.

4. Sample n^U polling stations at which undervoting irregularities are introduced. Set the share of polling stations with turnout discrepancies where unbalanced fraud is conducted to $S \in \{0, 0.02, 0.04, ..., 1\}$.

5. Iterate q times:

   (a) Add T_i^ϵ votes to those polling stations for which $D_i = 1$. Define the number of turned out voters in the main electoral race as $T_i = T_i^* + T^\epsilon$. For $n - n^U$ polling stations, the number of votes that is added (removed) from the winner is proportional to the winner's vote share V_i^*/T_i^* before undervoting discrepancies are introduced. For n^U polling stations, votes are added (removed) disproportionally. If $T_i^\epsilon > 0$, add a large share of votes to the winner and allocate the rest of the votes among the remaining candidates. If $T_i^\epsilon < 0$, a large share of votes is removed from the rest of the candidates and only few votes are removed from the winner.

   (b) For each polling station with undervoting discrepancies, construct u_i and p_i.

   (c) Compute the sum of the pointwise squared difference M_q between (p, u) from the main race in the simulated data and the empirical data.

Step 1 assures that the actual number and sizes of electorates for which data is generated is represented by the model across polling stations. Step 2 assures that the empirical dispersion in turnout, winner's votes and turnout discrepancies is represented by the model. After having performed this protocol q times for each possible fraud parameter $S \in \{0, 0.02, 0.04, 1\}$, the estimated portion of polling stations with undervoting irregularities that is supposed to be tainted is

$$\widehat{S} = \underset{S \in \{0, 0.02, 0.04, ..., 1\}}{\operatorname{argmin}} \operatorname{avg}(M). \tag{12}$$

Although the above protocol may seem complex, it has a very intuitive structure. It simply constructs synthetic data for two parallel electoral contests under the stochastic process defined in (7)-(9) using empirical input values and then constructs q different fraudulent elections for every

fraud parameter in S and computes the distance between each simulated election and the empirical data. The fraud parameter that—on average—leads to minimizing the distance between the empirical and simulated data is the estimate of fraud $\widehat{S}$, that is, the share of polling stations with undervoting irregularities where probabilistic fraud is assumed to have taken place. The model is non-parametric because in reverse-engineering the share of polling stations with undervoting discrepancies that are affected by fraud, it does not assume that the data generating process producing systematic alterations in the first place follows any particular functional form. Rather, the model flexibly finds the set of synthetic voting returns that are most similar to the empirical data.

There are two types of uncertainty associated with this method, namely fundamental and estimation uncertainty. Estimation uncertainty simply follows from the fact that for particular values of $\{\alpha^t, \beta^t, \alpha^v, \beta^v, \sigma\}$ constructed in Step 2, statistical sampling is performed throughout Steps 3-5. Uncertainty estimates that take into account estimation uncertainty simply estimate $\{\alpha^t, \beta^t, \alpha^v, \beta^v, \sigma\}$ from the data once using their maximum likelihood estimates, set q to any given value, and iterate over Steps 3-5 many times resulting in one estimate of the fraud parameter for each iteration. Uncertainty intervals can then straightforwardly be computed from the quantiles of $\widehat{S}$. Fundamental uncertainty refers to the fact that the parameters estimated in Step 2 that inform the sampling themselves are random variables and their true values are unknown. Taking into account uncertainty in the parameters is straightforward in a Bayesian setting, in which the parameters are first parameterized by conjugate prior distributions that put equal weight on the full range of plausible values and Steps 3-5 are then repeated for different posterior draws.[5,6]

In its technical setup, my model shares similarities with both the approach developed by Rozenas (2017) to detect rounding fraud from spikes in the density distribution of winner's votes and turnout and with the simulation model by Klimek et al. (2012) to identify ballot box stuffing from skewness, kurtosis and clusters within the bivariate distribution of turnout and votes for a single race. As Rozenas (2017), I rely on a stochastic model of elections defined as a sequence of binomial draws for the absolute numbers of turnout and winner's votes which are parameterized by beta distributions. Other than Rozenas, I don't define my model as a version of the Bayesian posterior predictive

[5]Details on prior distributions can be found in Appendix Section A2.

[6]This also means that the frequentist approach (incorporating estimation uncertainty) and Bayesian approach (incorporating estimation and fundamental uncertainty) to this method do not differ considerably in their computational efficiency, as both rely on the same number iterations of the algorithm. In a Bayesian setting, one simply first updates priors for three univariate probability distributions and executes the algorithm on different posterior draws rather than iteratively using fixed parameter values for $\{\alpha^t, \beta^t, \alpha^v, \beta^v, \sigma\}$ estimated from the data alone.

check. Rather, following Klimek et al. (2012), I reverse-engineer the level of fraud by iterating over a sequence of fraud parameters and choosing the one that minimizes a pre-specified distance metric. Methodologically, the method that I outline here hence combines features from different forensic methods that have been proposed in the literature. What makes the approach unique is exploiting the execution of parallel events, the focus on undervoting irregularities, and the novel quantity of interest that is ultimately retained: the share of polling stations with undervoting irregularities in which unbalanced fraud was conducted.

For executing the algorithm, the user has to define the relevant variables for constructing T^*, T^ϵ, V^*, specify the types of uncertainty that should be incorporated when constructing the fraud estimate, set q to an arbitrary large number and define the number of times the algorithm iterates over Steps 3-5. In case parameter uncertainty is supposed to be incorporated in Step 2, the user needs to define the set up for MCMC sampling. Appendix Section A3 displays an exemplary execution of the function for the Ecuadorian General Elections of 2017.

For the application of the model, a natural question arising is which of the parallel electoral contests should be used as the baseline election. Ultimately, this choice needs to be informed from substantive reasons and there is no statistical fix. The electoral contest that was allegedly fraudulent is set as the main race for which the fraud parameter is estimated, with the parallel contest working as the baseline election. In the case of more than two concurrent elections taking place, the algorithm needs to be repeated for each election pair of interest. The practice of election forensics goes hand in hand with qualitative and on-the-ground observations from political observers and institutionalized electoral observation and is no panacea providing quick answers without the substantive engagement of the researcher.

Applications

To illustrate the use of the proposed model, I apply it to a range of empirical elections and simulated electoral events for which the degree of fraud is known. The empirical data that I use are from the Ecuadorian General Elections of 2017 as well as the Ecuadorian Local Elections of 2019. Both elections have been subject to allegations of fraud, although the 2017 elections have been most controversial. All empirical data is available at the level of individual polling stations and comprise

up to $n = 39,322$ observations. Prior to the analysis of the empirical data, I exclude the rare polling stations which either registered less than $n = 100$ eligible voters or who report more turned out than eligible voters. I exclude the smallest polling stations because extreme percentages for turnout and winner's vote shares easily become artefacts of small electorates.[7] Polling stations with more turned out than eligible voters are excluded in order to facilitate the estimation of latent turnout and support rates through Beta distributions, which force success probabilities to range in the $[0, 1]$ interval.[8]

In addition, I apply the method to a range of simulated data sets that mimic two simultaneous electoral events. The simulated data follows the same stochastic process as outlined above. I apply the method to a fixed set of five artificial elections. In one of the elections, no fraud is introduced. In the remaining elections, the share of polling stations with undervoting irregularities at which unbalanced fraud is conducted varies between $\{0.2, 0.4, 0.6, 0.8\}$. The synthetic data that is constructed comprises $n = 20,000$ polling stations out of which $n^U = 1,000$ obtain discrepancies in turnout. All other parameters are set to similar values as found in the empirical data. All applications of the model that are presented incorporate both fumdamental and estimation uncertainty and thus report Bayesian credible intervals as uncertainty estimates.

Figure 4 re-constructs the plots on the relationship between undervoting irregularities and winner's vote shares for the General Elections 2017 that were introduced on simulated data in Figure 3. In the presidential contest, winner's vote shares refer to votes for the government and Correa-endorsed candidate (and ultimately elected president) Lenín Moreno. In the elections for the national and Andean assembly, winner's vote shares refer to the total share of votes that were cast for all candidates that ran for seats representing the government party MPAIS—*Movimiento Alianza País* (national assembly) or the electoral alliance between MPAIS and the Ecuadorian Socialist Party in the Andean elections. In the national referendum on civil cervants' and politicans' bank accounts in international tax havens, winner's vote shares refer to the share of votes cast for the government-endorsed option of accepting the reform that was posted. In all four electoral contests, the government-endorsed options received the largest overall vote share. Across all four elections,

[7] For an explicit test of voter rigging in small polling stations, see Jimenez, Hidalgo and Klimek (2017).
[8] Note: Empirical analyses have so far only be conducted for the 2017 elections. Results for the Local Elections 2019 are still to be included.

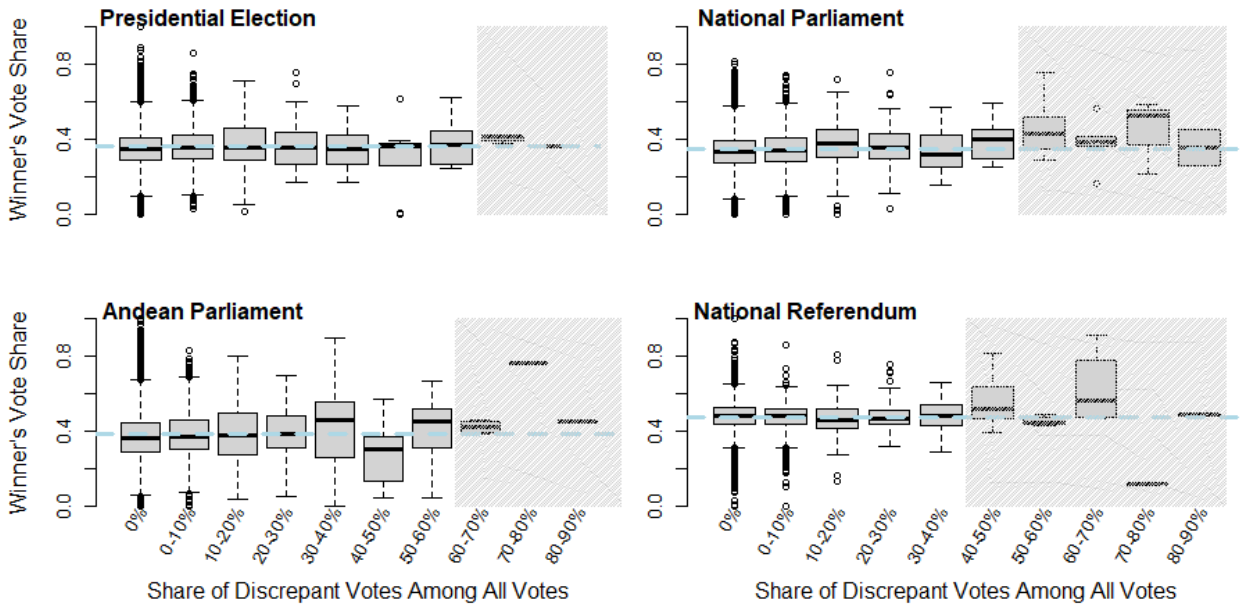


Figure 4: The relationship between undervoting irregularities and winner's vote shares (General Elections 2017). In the presidential contest, the winner is defined as Rafael Correa-endorsed candidate and election winner Lenín Moreno. In parliamentary elections, the winner's vote shares are defined as the total vote share of all candidates running for the government party Alianza País (national parliament) or the formed alliance with the Ecuadorian Socialist Party (Andean parliament). In the national referendum, the winner's vote share is the percentage of votes cast for the government-endorsed option.

the distribution of winner's vote shares varies homogeneously around the overall mean value for most extents of undervoting discrepancies. The vote shares of the election winner Lenín Moreno, the candidates of his associated party Alianza País, and the government-endorsed option in the national referendum administered on election day, however, are substantively skewed upwards in those polling stations that reported the most extreme values of undervoting. That is, vote shares for government-endorsed options are highest at those localities where irregularities are most extreme.

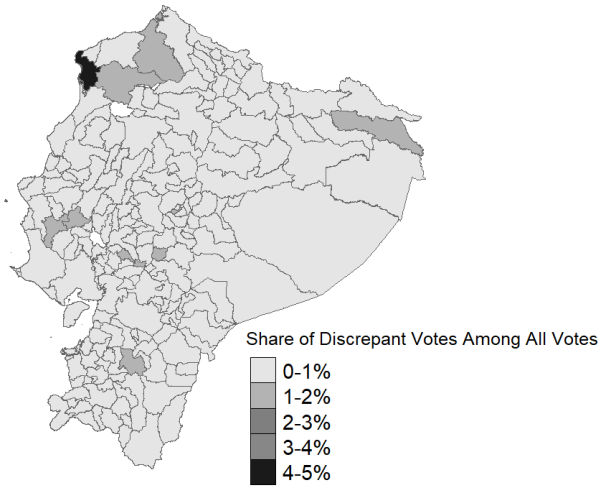
The results from the non-parametric simulation model for the empirical data and the batch of simulated elections are summarized in Table 1. Based on the model, undervoting irregularities in Ecuadorian voting returns of 2017 are well explained by unbalanced fraud approaches across the different electoral races. The share of polling stations with undervoting irregularities that is estimated to have witnessed unbalanced fraud ($\hat{S}$) ranges between 18% (elections for the Andean parliament) and 39% (presidential elections), which translates into an estimate of unbalanced fraud-

	IDs	IDs with Undervoting	Estimate ($\widehat{S}$)	95% Credible Interval
Ecuador Local Elections 2019				
Baseline: City Mayors				
Members of Parish Boards				
Rural Councilors				
Urban Councilors				
Provincial Prefects				
Ecuador General Elections 2017				
Baseline: Regional Parliaments				
Presidential Election	39,319	2,980	0.39	[0.24, 0.52]
National Parliament	39,319	2,340	0.21	[0.09, 0.33]
Andean Parliament	39,319	2,192	0.18	[0.09, 0.29]
National Referendum	39,319	2,748	0.35	[0.26, 0.44]
Simulated Elections				
0% Fraud	20,000	1,000	0	[0, 0.02]
20% Fraud	20,000	1,000	0.21	[0.10, 0.29]
40% Fraud	20,000	1,000	0.42	[0.30, 0.51]
60% Fraud	20,000	1,000	0.6	[0.51, 0.71]
80% Fraud	20,000	1,000	0.81	[0.69, 0.92]

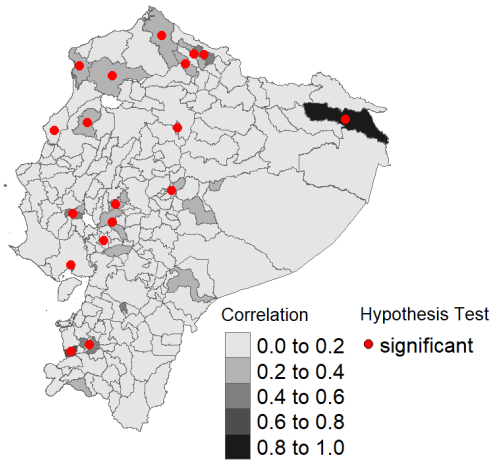
Table 1: Estimates of unbalanced election fraud. Non-parametric simulation models incorporate fundamental and estimation uncertainty and relie on 500 posterior draws in Step 2 and $q = 100$ iterations of Step 5. Column 'IDs' refers to the overall number of polling stations at which both races were administered. Column 'IDs with Undervoting' refers to the number of polling stations in which undervoting discrepancies are observed in relation to the baseline race. The fraud estimate $\hat{S}$ refers to the portion of polling stations with undervoting discrepancies at which unbalanced election fraud is supposed to be conducted. The last column presents Bayesian credible intervals.

ulent activity at 394 (or $1,162$) polling stations in the Andean (or presidential) elections. For the presidential elections, from the 95% credible intervals, we can say that with 95% probability this share lies between 24% and 52% of all polling stations with undervoting irregularities. This means that the distortions in group-specific distributions of winner's vote shares in Figure 4 are indicative of unbalanced fraud having interfered with the voting process at a substantial share of localities across the country.

Across the five artificial elections that have been simulated using different degrees of unbalanced fraud, true values for S are reliably reverse-engineered by the model, yielding confidence in the estimates that are constructed for the empirical data.



(a) Extent of undervoting irregularities.



(b) Correlations between extent of undervoting irregularities and Lenin Moreno's vote share.

Figure 5: Undervoting irregularities and their assocation to Lenin Moreno's vote share, Presidential Election 2017. **(a)** Map shows the average extent of undervoting irregularities across Ecuadorian cantons (excluding the Galápgos Islands) when comparing turnout in the presidential election to turnout for the baseline election of regional parliaments. **(b)** Map shows within-canton correlations between the extent of undervoting at a particular polling station and Lenin Moreno's vote share. The extent of undervoting at an individual polling station is defined as $u_i = \frac{T_i^{pres} - T_i^{reg}}{T_i^{pres}}$.

Figure 5 gives insight into the geographical distribution of undervoting irregularities and identifies the regional hotspots in which these are tied to unusually large vote shares for the winner in the presidential race that—according the simulation model—was most affected by interference. Finally, Table 2 reports a robustness test for the non-parametric simulation model across all polling stations of the country and presents linear multilevel regressions predicting winner's vote shares from the extent of undervoting u_i and a range of control variables. As can be seen from models M2, M7 and M8, only the two elections for which the estimated share of polling stations with unbalanced fraud is highest, unstandardized regression coefficients can be reliably distinguished from zero. This indicates that the non-parametric simulation model that I propose is more sensitive to detecting systematic irregularities than a parametric linear model. Figure 6 visualizes effects from Table 2.

	Dependent variable: Winner's vote share							
	Presidential Election		National Parliament		Andean Parliament		National Referendum	
	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)	(M7)	(M8)
Extent of Undervoting	0.013	0.020*	0.005	0.001	−0.005	−0.002	0.045***	0.046***
	(0.010)	(0.009)						
Closeness of the Electoral Race		0.001***		0.001***		0.001***		0.0001***
		(0.00001)		(0.00001)		(0.00002)		(0.00001)
Number of Eligible Voters		−0.0002***		−0.0001***		−0.0002***		0.00000
		(0.00001)		(0.00001)		(0.00001)		(0.00001)
Percentage Turnout		0.056***		0.061***		0.067***		0.028***
		(0.004)		(0.004)		(0.005)		(0.004)
Percentage Null Votes		−0.120***		0.225***		0.249***		−0.186***
		(0.011)		(0.012)		(0.015)		(0.010)
Percentage Blank Votes		0.089***		0.414***		0.518***		−0.258***
		(0.016)		(0.018)		(0.022)		(0.015)
Constant	0.360***	0.333***	0.342***	0.266***	0.382***	0.300***	0.468***	0.465***
	(0.006)	(0.007)	(0.006)	(0.007)	(0.007)	(0.008)	(0.004)	(0.005)
N Polling Stations (N Cantons)	39,319 (251)	39,319 (251)	39,319 (251)	39,319 (251)	39,319 (251)	39,319 (251)	39,319 (251)	39,319 (251)
ICC	0.68	0.61	0.68	0.58	0.59	0.56	0.49	0.50
R Squared	0.68	0.73	0.61	0.64	0.59	0.63	0.49	0.51

Table 2: The relation between undervoting irregularities and winner's vote shares, General Elections 2017. *Note: Table presents unstandardized coefficients from linear multilevel regression models with random intercepts across 251 cantons fitted with maximum likelihood estimation. Standard errors are reported in parentheses. *p<0.05; **p<0.01; ***p<0.001.*

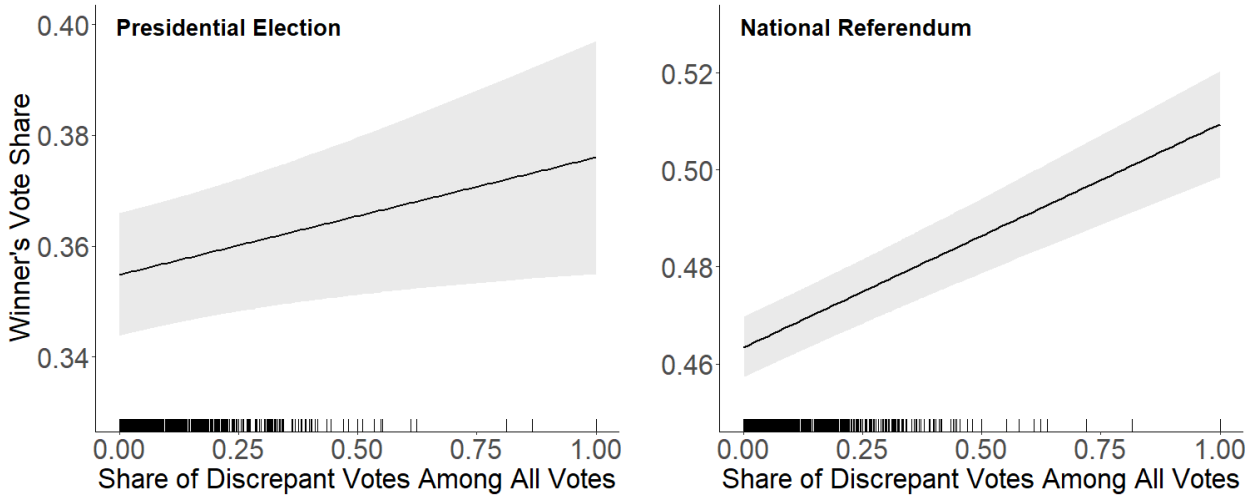


Figure 6: The relationship between undervoting irregularities and winner's vote shares, General Elections 2017. The figures plot expected values simulated under models M2 and M8 using 10,000 draws from a multivariate normal distribution defined by the vector of parameter estimates and their covariance matrix. All control variables reported in Table 2 are held constant at their mean values. Shaded regions visualize the 2.5% and 97.5% quantiles of the simulated expected values.

Alternative Explanations

As outlined in Section 3 of this paper, if discrepancies in turnout across multiple electoral races are produced at random due to the miscount or loss of votes, no statistical relationship is expected between the extent of undervoting and the winner's vote share. Through the application of the non-parametric simulation model and additional statistical analyses, the former section outlined that the empirical patterns that are inherent to Ecuadorian voting returns are indicative of non-random processes producing undervoting irregularities and are well explained by the mechanism of unbalanced fraud.

However, not all mechanisms that are non-random equal fraudulent activity. Importantly, there is room for alternative explanations that do not evoke fraud which would lead to similar empirical patterns. A potential explanation for a substantive relationship between the extent of undervoting irregularities and vote shares of government-endorsed options stems from the empirical finding that votes for the populist left in Latin America have been found to be correlated to unemployment levels and economic hardship (Queirolo 2013). It is straightforward to assume that (i) political knowledge is a major determinant of voters not casting votes in all electoral contests for which they are eligible and (ii) that political knowledge is lower among the part of the population for which

poverty is highest. Especially poorer voters, however, might get mobilized by leftist presidents and an electoral campaign focusing on a platform centered around social spending and re-distributional policies. Under this mechanism, regions in which government-endorsed vote shares are highest are also expected to witness the largest amounts of undervoting irregularities, leading to similar empirical pictures are summarized above. In general, it is unclear whether the proposed non-parametric simulation model reverse-engineers unbalanced fraud or picks up on this alternative mechanism that— if in place—produces turnout discrepancies in a systematic way with similar empirical patterns.

Future versions of this paper will need to empirically trace observable implications of this alternative theory that does not evoke fraud to explain the empirical patterns summarized in Figures 4 and 6 and that get picked up by the simulation model in Table 1:

- *On the individual level:* Does the link between economic hardship, reduced levels of political knowledge, and leftist voting hold for Ecuadorian election survey data collected in 2017?
- *On the aggregate level:* Are the cantons that report significant correlations in Figure 5(b) those with lowest levels of economic development?

Conclusion

This paper exploited the execution of several concurrent electoral contests on election day for the statistical detection of election fraud. I presented the country case of Ecuador and showed that the execution of simultaneous electoral events can give rise to a phenomenon called 'undervoting irregularities', which occur if the same polling stations document different numbers of turned out voters for different electoral contests. A series of logically equivalent transformations showed that if undervoting irregularities are produced by a random process, all candidates are affected equally by these and there is no statistical expectation of a covariance between the extent of undervoting and the winner's vote shares across localities. Next to describing undervoting irregularities under random processes stemming from limited capacities of low-level election officials, I introduced the systematic mechanism of unbalanced fraud which occurs if protagonists of fraud fail to interfere into multiple electoral races to equal extents.

The paper proposed a non-parametric simulation method to detect unbalanced fraud approaches from undervoting irregularities and their relation to winner's vote shares. Using the method, practitioners of election forensics can estimate the share of polling stations with undervoting at which unbalanced fraud has been perpetrated and quantify the uncertainty of estimates under different statistical paradigms.

The method that I proposed only focuses on one very specific kind of fraud, namely the unequal manipulation between a main race of interest and a baseline election. This is not equal to estimating the overall degree of fraud that might be inherent to published electoral data as several different mechanisms of systematic manipulation that the model is not designed to pick up might be at place. Vice versa, the method does not assume that the baseline election itself is actually fraud-free. Rather, I present a statistical approach to reverse-engineer the degree of unequal intervention across multiple races.

Lastly, while the empirical patterns that are inherent to Ecuadorian General Elections of 2017 are well explained by unbalanced fraud approaches, it is important to note that there are alternative mechanisms that do not center around any kind of fraudulent activity which can produce similar empirical pictures. Practitioners of the method need to pay close attention to these alternative mechanisms. Ultimately, only careful data analyses providing robust evidence against the existence of alternative mechanisms that go along with the estimates from the non-parametric simulation model speak in favor of systematic manipulation.

References

Beber, Bernd and Alexandra Scacco. 2012. "What the Numbers Say: A Digit-Based Test for Election Fraud." *Political Analysis* 20(2):211–234.

Callen, Michael and James D. Long. 2015. *Institutional corruption and election fraud: Evidence from a field experiment in Afghanistan.* Vol. 105.

Cantú, Francisco. 2014. "Identifying Irregularities in Mexican Local Elections." *American Journal of Political Science* 58(4):936–951.

Cantú, Francisco. 2019. "The Fingerprints of Fraud: Evidence From Mexico's 1988 Presidential Election." *American Political Science Review* 113(3):710–726.

Carter Center. 2007. "Final Report on Ecuadors September 30, 2007, Constituent Assembly Elections." *Election Observation Report* .

de la Torre, Carlos. 2015. *De Velasco a Correa: Insurrecciones, populismos y elecciones en Ecuador, 1944-2013.* Quito, Ecuador: Corporación Editora Nacional - Unversidad Andina Simón Bolívar.

EU, European Union. 2002. "Final Report on the Presidential and Parliamentary Elections, Ecuador." *Election Observation Report* .

EU, European Union. 2009. "Final Report on Presidential and Parliamentary Elections 26 April, 2009, Ecuador." *Election Observation Report* .

Haerpfer, C., R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin and B. Puranen. 2022. *World Values Survey: Round Seven - Country-Pooled Datafile Version 3.0.* Madrid, Spain and Vienna, Austria: JD Systems Institute and WVSA Secretariat.

IRI, International Republican Institute. 2003. "Republic of Ecuador, National Elections October 20 November 24, 2002 - Report and Recommendations." *Election Observation Report* .

Jimenez, Raúl, Manuel Hidalgo and Peter Klimek. 2017. "Testing for voter rigging in small polling stations." *Science Advances* 3(6):1–7.

Klimek, Peter, Yuri Yegorov, Rudolf Hanel and Stefan Thurner. 2012. "Statistical Detection of Systematic Election Irregularities." *PNAS* 109(41):16469–16473.

Kobak, Dmitry, Sergey Shpilkin and Maxim S. Pshenichnikov. 2016. "Integer percentages as electoral falsification fingerprints." *Annals of Applied Statistics* 10(1):54–73.

Mebane, Walter R. 2008. Election Forensics: The Second-Digit Benford's Law Test and Recent American Presidential Elections. In *Election Fraud: Detecting and Deterring Electoral Manipulation*, ed. R. Michael Alvarez, Thad E. Hall and Susan Hyde. Washington, D.C.: Brookings Institution Press pp. 162–181.

Medzihorsky, Juraj. 2015. "Election Fraud: A Latent Class Framework for Digit-Based Tests." *Political Analysis* 23:506–517.

Myagkov, Mikhail, Peter C. Ordeshook and Dimitri Shakin. 2009. *The Forensics of Election Fraud: Russia and Ukraine.* Cambridge: Cambridge University Press.

Norris, Pippa. 2014. *Why Electoral Integrity Matters.* Cambridge: Cambridge University Press.

OAS, Organization of American States. 1998. "Electoral Observation in Ecuador 1998." *Election Observation Report* .

OAS, Organization of American States. 2006. "Informe de la misión de observación electoral de la OEA en la república del Ecuador." *Election Observation Report* .

OAS, Organization of American States. 2008. "Informe de la misión de observación electoral de la OEA en la república del Ecuador." *Election Observation Report* .

Queirolo, Rosario. 2013. *The Success of the Left in Latin America: Untainted Parties, Market Reforms, and Voting Behavior.* Notre Dame: University of Notre Dame Press.

Rozenas, Arturas. 2017. "Detecting Election Fraud from Irregularities in Vote-Share Distributions." *Political Analysis* 25:41–56.

Supplementary Material for Detecting Unbalanced Election Fraud Approaches From Undervoting Irregularities

Lion Behrens
University of Mannheim
behrens@uni-mannheim.de

June 17, 2022

Contents

1 The Distribution of Undervoting Irregularities

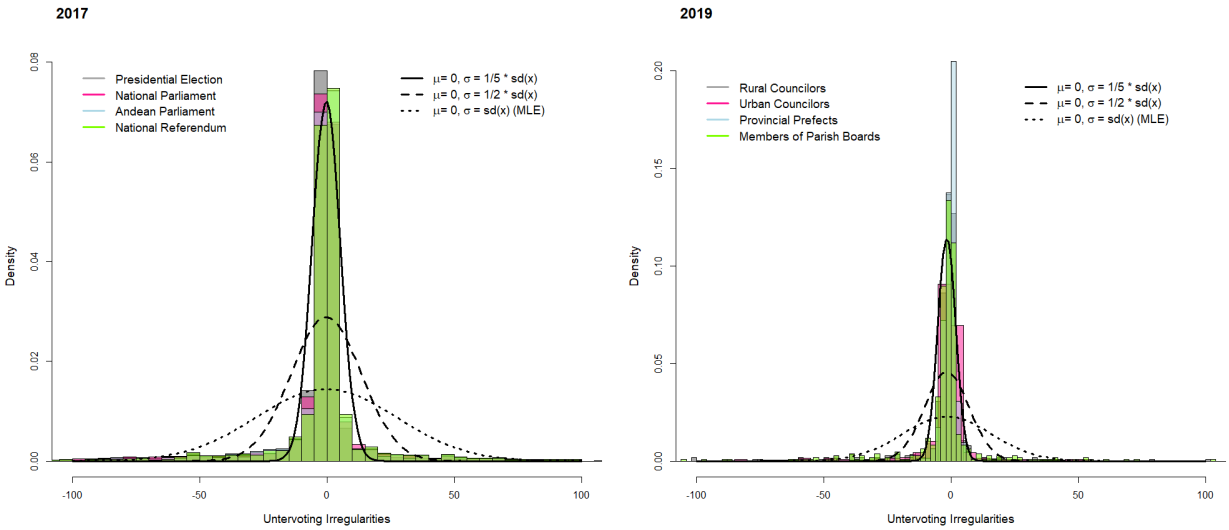


Figure 1: The distribution of undervoting irregularities in Ecuadorian elections. Left panel: General Elections 2017. Right panel: Local Elections 2019. Histograms visualize the discrepancies in absolute numbers of documented turned out voters between the election of state-level members of parliament (2017) and city mayors (2019) and four concurrent elections. Black lines depict normal distributions scaled by different dispersion parameters that are estimated from the data.

Observed turnout levels $\mathcal{T}_i$ can be decomposed as

$$\mathcal{T}_i = \mathcal{T}_i^* + \mathcal{T}_i^\epsilon i, \tag{1}$$

where $\mathcal{T}_i^* \in [0, N_i]$ is the true number of total votes cast and $\mathcal{T}_i^\epsilon \in [0, N_i]$ is the absolute number of votes that has been added (removed) by error or fraud. Across all polling stations $i \in \{1, ..., n\}$, turnout discrepancies $\mathcal{T}_i^\epsilon$ are distributed as

$$\mathcal{T}_i^\epsilon \sim \mathcal{N}(\mu, \sigma^2). \tag{2}$$

Figure 1 shows that the normality assumption fits the data well. Undervoting irregularities are dispersed around a mean of $\mu = 0$. The maximum likelihood estimate of the dispersion parameter is given by $\sigma = \sqrt{\frac{\sum_{i=1}^{n} \mathcal{T}_i^\epsilon - \mu}{n}}$ and provides a poor fit to the kurtosis of the distributions. Estimating the dispersion parameter as

$$\sigma = \frac{1}{5} * \sqrt{\frac{\sum_{i=1}^{n} \mathcal{T}_i^\epsilon - \mu}{n}} \tag{3}$$

provides a close fit to undervoting irregularities (i) across different elections (ii) and years.

2 Prior Distributions for Parameter Estimation

In order to sample from the distributions outlined in Equations (7)-(9), parameters $\{\alpha^t, \beta^t, \alpha^v, \beta^v, \sigma\}$ can either be estimated using maximum likelihood estimation from the empirical data that is being used. Alternatively, to incorporate fundamental uncertainty of the parameters and iterate the algorithm across a range of different parameter values, parameters can be parameterized by prior distributions. The execution of the non-parametric simulation model is then repeated once for each posterior sample. The following prior distributions underlie the execution of the model:

$$\begin{aligned}
\sigma &\sim \text{InvGamma}(0.001, 0.001) \\
\alpha^t &\sim \text{N}(0, 1000) \\
\beta^t &\sim \text{N}(0, 1000) \\
\alpha^v &\sim \text{N}(0, 1000) \\
\beta^v &\sim \text{N}(0, 1000)
\end{aligned}$$

3 R Code for Executing the Non-Parametric Simulation Model

The code snippet below documents an exemplary execution of the non-parametric simulation model by the user for data from the Ecuadorian General Elections 2017. The election for members of regional parliaments (*asambleístas provinciales*) is set as the baseline electoral contest. Estimated is the share of polling stations with undervoting at which unbalanced fraud between the election of interest and the baseline electoral contest is perpetrated. Function calls incorporate fundamental and estimation uncertainty and output estimated shares together with 95% credible intervals.

```
library(EnvStats)
library(fields)
library(foreign)
library(rstan)

# load data at polling station level, General Elections 2017
load("actas17.Rdata")

# delete polling stations with <100 eligible voters
actas17 <- actas17[-which(actas17$ELECTORES_REGISTRO_pres<100),]

# run model, presidential election
actas17 <- actas17[-which(actas17$turnout_pres>1),] # exclude if turnout > 1
ecu17_pres <-
  est_fraud(eligible = actas17$ELECTORES_REGISTRO_pres,
            turnout_main = actas17$SUFRAGANTES_pres,
            turnout_baseline = actas17$SUFRAGANTES_asam_prov,
            winner_main = actas17$MORENO_pres,
            uncertainty = c("fundamental", "estimation"),
```

```
            n_iter = 100,
            n_postdraws = 500,
            n_burnin = 400,
            seed = 12345
            )

# run model, national parliament election
actas17 <- actas17[-which(actas17$turnout_nac>1),] # exclude if turnout > 1
ecu17_nac <-
  est_fraud(eligible = actas17$ELECTORES_REGISTRO_asam_nac,
            turnout_main = actas17$SUFRAGANTES_asam_nac,
            turnout_baseline = actas17$SUFRAGANTES_asam_prov,
            winnershare_main = actas17$winnershare_asam_nac,
            uncertainty = c("fundamental", "estimation"),
            n_iter = 100,
            n_postdraws = 500,
            n_burnin = 400,
            seed = 12345
            )

# run model, Andean parliament election
actas17 <- actas17[-which(actas17$turnout_andean>1),] # exclude if turnout > 1
ecu17_andean <-
  est_fraud(eligible = actas17$ELECTORES_REGISTRO_andino,
            turnout_main = actas17$SUFRAGANTES_andino,
            turnout_baseline = actas17$SUFRAGANTES_asam_prov,
            winnershare_main = actas17$winnershare_andino,
            uncertainty = c("fundamental", "estimation"),
            n_iter = 100,
            n_postdraws = 500,
            n_burnin = 400,
            seed = 12345
            )

# run model, national referendum
actas17 <- actas17[-which(actas17$turnout_referend>1),] # exclude if turnout > 1
ecu17_referendum <-
  est_fraud(eligible = actas17$ELECTORES_REGISTRO_consulta,
            turnout_main = actas17$SUFRAGANTES_consulta,
            turnout_baseline = actas17$SUFRAGANTES_asam_prov,
            winner_main = actas17$Si_consulta,
            uncertainty = c("fundamental", "estimation"),
            n_iter = 100,
            n_postdraws = 500,
            n_burnin = 400,
            seed = 12345
            )
```