Anomalies and Frauds in the Korea 2020 Parliamentary Election, SMD and PR Voting with Comparison to 2016 SMD: Comment

Youn Baek*

June 1, 2020

Abstract

Using `eforensics`, Mebane (2020a) documents anomalies in the voting patterns of the 2020 Korea Parliamentary Election that may indicate the existence of large-scale election frauds. The estimated probability of election frauds is 6.7%-7.3. This paper proposes a parsimonious and tractable approach that identifies correctly specified models that are required by `eforensics`. The estimated probability of election frauds in the correctly specified models is 1.3%-1.8%. Mebane (2020a)'s model misspecification dervies from misperception about turnout rates for one particular voting method—*prevote*—as close to 100%, while the national turnout rates for *prevoting* is 25.52%. We test a set of different specifications to investigate whether potential pitfalls of our alternative approach undermine our results.

*Ph.D. Student, NYU Stern School of Business. syb262@stern.nyu.edu

†I thank Walter Mebane for helpful discussions and sharing replication codes; Kandol Lee and Jong Hee Park for comments; Alistair Barton and Jonathan Becker for helping me understand Canadian and German elections. All errors are entirely mine. This manuscript is preliminary and please do not circulate it without permission.

1 Introduction

This paper does not aim to provide any evidence for or against election frauds in the 2020 Korea Parliamentory Election. The goal of this paper is to do replication exercises of Mebane (2020a) by running `eforensics` package with an alternative set of specifications. We restrict our attention to the analysis from `eforensics` in the context of 2020 South Korea election[1]. When we calculate more reasonable inputs for `eforensics` model, we obtain results that are substantially different from those in Mebane (2020a). To understand why we see this contradiction, we briefly explain a specific way of casting a vote—prevoting—and show that understanding this voting process is crucial to identifying correctly specified models.

In April 2020, the South Korean legislative election was held. Out of 43,994,247 eligible voters in South Korea, 29,126,396 voters participated in the parliamentary election(Korea National Election Commission, 2020) [2]. 17,247,978 voters showed up for the election-day voting and 11,742,677 voters opted in for the prevote[3].

By analyzing polling station data of the Korean election, Mebane (2020a) explores the voting patterns in the single-member district(SMD) seats and proportional representation(PR) seats.

With SMD data, he estimates the probability of election fraud under two different scenarios. The first one assumes that the incumbent Democratic Party benefited from frauds(Democratic Party specification), and the second one assumes that whoever won the seat in each constituency benefited from frauds(constituency leader specification).

Under Democratic Party specification, he demonstrates that election frauds happened in 28.7%, 2.43%, 0.67%, and 0.61% of prevote polling stations, election day polling stations, voting by post units, and abroad election polling stations respectively. In this scenario 1,418,079 votes are fraudulent where 1,056,462 votes are manufactured and 361,617 votes are stolen by the candidate of Democratic Party. According to this statistical model, the

[1]For general discussions about 2BL method, for instance, see Deckert, Myagkov and Ordeshook (2011), Klimek, Yegorov, Hanel and Thurner (2012), and Mebane (2011)

[2]In Mebane (2020a)'s analysis these numbers are 43,961,157 and 28,738,468. This discrepancy might have happened for a variety of reasons, but it would not matter for the analysis as long as they are not too far off.

[3]The meaning of prevote will be explained in detail in the next section.

probability of no fraud is 92.7%; incremental fraud is 6.15%; extreme fraud is 0.51%.

Under constituency leader specification, the model reports that 25%, 2.04%, 1.52%, and 0% of prevote polling statios, election day polling stations, voting by post units, and abroad election polling stations votes are fraudulent. In this scenario 1,234,217 votes are fraudulent where 961,296 votes are manufactured and 272,921 votes are stolen by the winners of each constituencies. The estimated probability of no fraud is 93.3%; incremental fraud is 6.51%; extreme fraud is 0.78%.

Among 253 constituencies, there are 22 constituencies where the election outcome would have changed had it not been for the election fraud detected from the `eforensics` model. 15 of them were from the Democratic Party and the remaining 7 were from Future Integration Party.

With PR data, the Mebane (2020a) estimates election frauds under two different scenarios. One of them assumes that Future Korea Party benefits from the fraud and the other one assumes that Platfor Party benefited[4].

In the first scenario where Future Korea Party benefited, 16.6%, 11.6%, 0.40%, and 0.24% of prevote, election-day votes, abroad voting, and voting post-election day votes are fraudulent. 636,694 votes are fraudulent where 448,143 are manufactured and 188,551 are stolen. This model indicates that 6.7% of Future Korea Party's proportional representation legislators are fraudulently elected. Based on this result, the estimated probability of no fraud is 95.5%; incremental fraud is 4.44%; extreme fraud is 0.02%.

In the second scenario where Platform Party benefited, they are 18.7%, 26.0%, 0.32%, and 0%. Here 571,665 votes are fraudulent where 366,390 are manufactured and 205,275 are stolen. This model indicates that 6.1% of Platform Party's proportional representation legislators are fraudulently elected. Based on this result, the estimated probability of no fraud is 96.2%; incremental fraud is 3.82%; extreme fraud is 0.0058%. Based on several supplementary analysis, Mebane (2020a) suggests that PR results are more likely to have

[4]Future Korea Party and Platform Party are subordinate political parties of Future Integration Party and Democratic Party. The electoral reform that took effect in 2020 gives disadvantages to the parties with a large number of SMD seats in the PR election to provide small and minor political parties with more opportunities to take seats in the proportional representation. In response to this reform, the two major political parties, Democratic Party and Future Integration Party, created subordinate political parties controlled by them to bypass this electoral regulation.

been driven by strategic behaviors of voters.

For simplicity, this paper will restrict the focus on the analysis of SMD results and probability estimation of election frauds in the model. In general, a statistical analysis can be erroneous for three different reasons. First, data are faulty. Second, models are unrealistic. Third, implementation is wrong. In order to validate Mebane (2020a)'s analysis, we have to check all three of them. However, we do not aim to do the first two tasks in this paper. Instead, we focus on the implementation validity.

There are two reasons why this paper focuses on the implementation validity. First, Mebane acknowledges the lack of domain knowledge about the electoral system, the party system, and voting behaviors in South Korean democracy. For one thing, Mebane (2020a) translates Korean language into English using Google Chrome when implementing his analysis.

Second, South Korea's 2020 election was unprecedented in many aspects. The prevote turnout was the highest in record and the electoral system was changed. Major parties enlisted their satellite parties in the PR list to gain maximum votes from the new electoral rule. All of these factors confound analysis that tries to uncover hidden out of election outcomes.

We show that the results from `eforensics` change when we plug values required by the `eforensics` model. In the following sections, we explain what *prevote* is, and some elementary concepts of Mebane (2019b) 's model to describe a way to estimate correctly specified models that eliminate statistical artifacts engendering spurious results.

2 Institutional Background

2.1 What is Prevote?

The key to grasping the reasons why we obtain contrasting results is to understand one way of casting a vote in the Korean election—prevoting[5]. In South Korea, voters are allowed

[5] `https://www.nec.go.kr/portal/bbs/view/B0000357/7153.do?menuNo=200561&searchYear=&searchMonth=&searchWrd=&searchCnd=&viewType=&pageIndex=1`

to prevote instead of voting on the election day[6]. Prevoting is held on the 5th and 4th days before election day in polling stations across the country. On the prevoting days voters bring their ID card to the polling station and they can cast their votes without any further restrictions—no preregistration is required. Indeed, every eligible voter is automatically preregistered for prevoting. If the voter wants to participate in prevoting, she can just shows up at the polling station with her ID. If she wants to wait until election day, she has to do nothing to opt out of prevoting. She just needs to wait until election day and cast her vote.

One advantage of prevoting compared over election day voting is that people can participate in the election even when they are remotely located from their constituencies. When people go to the prevote polling station on the prevoting day, the voting machine prints out customized ballots for each person based on the address of the voters. Voters whose official addresses are located in constituency i can cast their ballots in constituency $j \neq i$ in the prevote. Prevote can promote electoral participation by reducing the transportation costs for people who would not otherwise have been able to return to their constituencies that are far from their current residence to cast their votes. If voters opt in for prevoting, they can cast their ballots anywhere in the country without preregistration or any further restrictions. This is possible because the list of every eligible voter in South Korea is stored in a centralized system that could confirm each voters' identity and whether the voters have already cast their votes or not[7]. Hence, prevoting is conceptually different from absentee voting or abroad voting where a voter does not have the option of on-site same-day registration.

Prevoting is also distinguished from early voting in New York State. In New York State, early voting was signed into law in 2019 and took effect in January 2020 for the first time in history to allow voters to cast their ballots ahead of election day. However, the same-day on-site registration for early voting is not possible yet. Preregistration is required for early

[6]Voting-in-advance would also be a good term for capturing the meaning of this voting method. Hereafter I use prevote to follow Mebane (2020a)'s terminology.

[7]This is similar to voting at the advance poll in Canadian election. In Canada, voters are allowed to cast their ballots 10th-7th days before election day without additional restrictions relative to the election day voting. It is not exactly the same as South Korea because voters can cast their ballots outside of their constituencies in South Korea. `https://www.elections.ca/content.aspx?section=vot&dir=vote&document=index&lang=e` On the other hand, there is no such voting method in Germany.

voting in New York. As of June 2020, the New York State Constitution prevents voters from casting their ballots unless they register for voting at least ten days before election day(New York State, 2019). If same-day registration for early voting comes into effect in the future, early voting will technically be similar to prevoting in South Korea. Even then, South Korean prevoting is free from geographical restriction as long as a voter is within South Korea while voters in the US cannot participating outside of their own constituencies. Additionally, it is unclear whether the Federal Election Commission in the US will publish its election data on early voting the way Korea National Election Commission does on prevoting. The data management issue will be discussed further in the section 2.3.

The number of voters who are going to participate in the prevote is unknown until the prevote is finished. Therefore, it is vital to note that the number of "eligible" voters in the prevote are the total number of eligible voters in each constituencies. For every other voter who did not participate in the prevote nor register for absentee/abroad voting is assigned to a particular polling station within their constituency. On the official election day, voters cannot vote unless they go to designated polling stations within each constituencies.

To sum up, there are two days for prevoting and one day for election day voting. Technically, there are less restrictions for prevoting in the sense that prevoters can cast their ballots in polling stations that are outside of their constituencies without any preregistration that informs the election committee of their willingness to participate in prevoting prior to prevoting day.

2.2 Some Notations in Mebane (2019b)

This subsection introduces elementary concepts of Mebane (2019b) that helps understand the reasons why we get contrasting results. Some of the primary inputs in `eforensics` model are the number of eligible voters N_i, the number of voters valide votes V_i, and the turnout proportion $t_i = \frac{V_i}{N_i}$ for each aggregation unit i. The model in the `eforensics` package estimates the predicted turnout proprtion τ_i, which is considered the true turnout proportion. For each i, the fraction of manufactured votes is denoted as ι_i^M.

Now suppose there was an 'incremental fraud' as opposed to an 'extreme fraud' where

the meaning of each term is defined in Mebane (2019b). Under an incremental fraud,

$$\mathbb{E}(1 - t_i) = (1 - \tau_i)(1 - \iota_i^M) \tag{1}$$

Thus if the observed turnout proportion is higher than the predicted value so that $t_i > \tau_i$, then the potential election frauds could be inferred from the proportion of manufactured votes $\iota_i^M > 0$. By analyzing polling station-level data, Mebane (2020a) estimates τ_i based on the fixed effects of prevote, voting post, abroad, and mobile voting[8] of each polling station and the fixed of each constituency, and then compare these estimates τ_i with observed turnout proportion t_i to detect manufactured votes.

2.3 Korean Election Data

The Republic of Korea National Election Commission provides the election statistics on their webpage[9]. One feature that could be confusing for those who are not familiar with this data or Korean electoral institutions is the column "the number of voters" in the data. For *election day* outcomes, this value refers to the actual number of eligible voters who are specifically designated to that particular polling station.

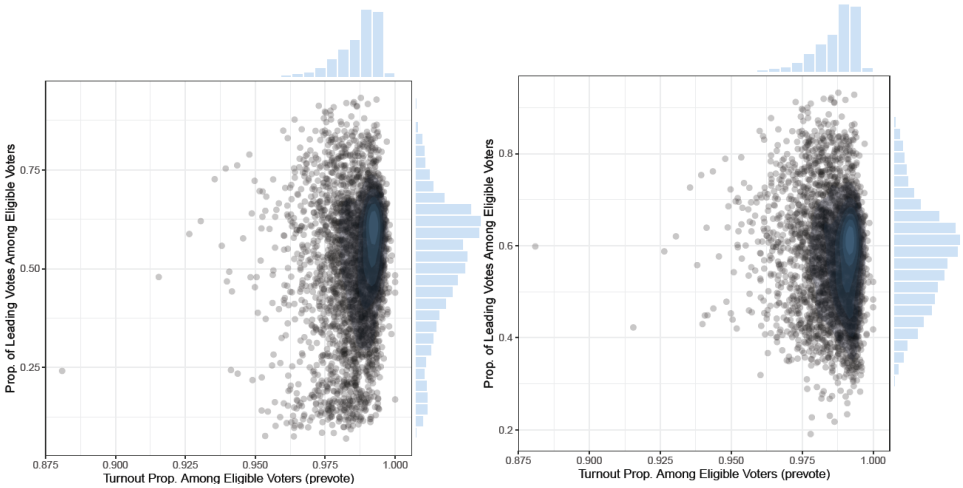
On the other hand, for every *prevote* polling station this "the number of voters" column denotes the number of people who opt in for the prevote instead of other voting methods. It does not mean the total number of predetermined people who are eligible to cast their votes in the prevote. Even if they are placed in the same column, the interpretation of each values of "number of voters" is inconsistent across polling stations depending on whether they are prevote polling stations or election day polling stations.

Right next to the column indicating the "number of voters" is "the number of votes." The number of "valid votes" can be calculated by subtracting invalid votes from the number of votes. These values have a consistent interpretation .

For every *election day* polling station, the fraction of "valid votes" over "the number of voters" does mean the observed turnout proportion that should be put in place in Mebane (2019b)'s model. In mathematical notation, this is essentially $t_i = \frac{V_i}{N_i}$. This makes sense

[8] Mebane (2020a) calls it disabledship status, but it is more appropriate to call it mobile voting.

[9] `http://info.nec.go.kr/`



(a) Democratic Party specification (b) Constituency leader specification

Figure 1: Prevote data plots by Mebane (2020a)

Notes: Horizontal and vertical axis indicate the turnout proportion among eligible voters, and the vote rate of the (a)Democratic Party and (b)constituency winner candidates. The mean of prevote turnout rates across 3,802 prevote polling stations in Mebane (2020a) is 98.62%.

because for the election day voting, the number of eligible voters for each polling stations are predetermined before election day and the value of N_i can be fixed.

On the contrary, for every *prevote* polling station, this fraction denotes the number of valid votes over the sum of the number of valid votes and the number of invalid votes. Let δ_i denote the number of invalid votes in prevote polling station. Then the fraction of the column of the number of "valid votes" over the column of the "number of voters" could be expressed as $\frac{V_i}{V_i+\delta_i}$. This is not equal to $t_i := \frac{V_i}{N_i}$ unless every single eligible voter in constituency i is enforced to participate in prevote and is forbidden from election day voting. This is not what happened in the South Korean election.

Mebane (2020a) uses this value, $\frac{V_i}{V_i+\delta_i}$, as an observed turnout proportion of prevote so that $t_i = \frac{V_i}{V_i+\delta_i}$ for prevote polling stations. In any of the 253 constituencies, $N_i = V_i + \delta_i$ does not hold. The scatter plot in Mebane (2020a) shows that most of the observed turnout proportion of prevote are concentrated in 97%-100% area. This is incorrect by any measure. The national prevote turnout rate is 25.52%(Yonhap News, 2020). It is obtained by dividing the number of prevoters by the number of total eligible voters except for those who preregistered for abroad voting or mobile voting.

In a personal email exchange, Mebane (2020b) says that this misunderstanding of raw data is not problematic at all since he controls for an indicator variable which is equal to one if the observation is a polling station of the prevote. There are two reasons one could be skeptical of this claim even before doing replications.

First, the prevote turnout rates in Mebane (2020a) is not *turnout rates*, an input that is required by Mebane (2019b)'s `eforensics` model. An indicator variable account for for qualitative differences between groups within a dataset(Wooldridge, 2016). An indicator variable is not a tool that fixes statistical artifacts originating from the confusion about raw data. The definition of turnout proportion is the fraction of voters who showed up to cast their ballots among eligible voters. For prevote, the eligible voters are every eligible voter in each constituencies. It is not equal to the number of people who showed up to cast their ballots on prevoting days. What he uses as the values for prevote turnout proportion in his paper is not what `eforensics` model requires as inputs for turnout proportion since $t_i \neq \frac{V_i}{V_i+\delta_i}$.

This misconception does have a great impact on estimating the number of "manufactured votes." The estimated coefficients of Table 1 in Mebane (2020a) display that the intercept term if 0.766 and the fixed effect of prevote is 1.10. The constituency fixed effects are much smaller than these values. When we plug these parameters into the logistics function, the estimated turnout proportion, which is considered true turnout proportion in `eforensics` model, is around $\frac{1}{1+\exp(-0.766-1.1)} \approx 86.6\%$. Given that most of the prevote polling stations exhibit observed turnout proportion of close to 100%, and based on the way his model detects manufactured votes, it is quite clear that the turnout proportions of prevote polling stations that are close to one are playing crucial roles in counting "manufactured votes." Since around one fifths of entire polling stations are prevote polling stations that show turnout proportion of 100% while others show much lower turnout rates, it is no surprise to find that polling stations with turnout proportion of 100% are considered rigged in `eforensics` model. If this problem can be easily fixed by merely controlling for an indicator variable, on the other hand, then it means that `eforensics` model is not useful for detecting election frauds such as ballot box stuffing(Klimek, Yegorov, Hanel and Thurner,

2012)[10].

One could still argue that the overall result that election frauds are detected would not change under different specifications. Even when this is true, secondly, this approach evidently ignores the crucial information that could be inferred from data. The prevote turnout proportions were 25.52% for Busan Metropolitan City, 32.37% for Sejong Special Self-Governing City, and 24.65% for Jeju-do to name a few(Yonhap News, 2020)[11]. It goes without saying that these *correct* turnout proportions are calculated by dividing the number of prevote participants by the total number of eligible voters in each constituencies. If we equalize an array of the heterogeneous turnout proportions into a narrow set of concentrated values ranging from 97%-100%, then it is very unlikely to detect potential election frauds in a correct way.

Note that this is not to say that there should not be an election fraud when we use proper prevote turnout proportions. Using correct values of prevote turnout proportion will help us identify the constituencies where the winners would have changed in a counterfactual scenario, as opposed to detecting the wrong places where frauds did not occur.

3 Replications

3.1 Panel aggregation approach

To analyze election data within `eforensics` model we need sensible values that indicate the turnout proportion of every unit of analysis. Among many ways of operationalizing this idea, the baseline approach of this paper is to aggregate the polling station data into a panel data of (253 constituencies $\times$ (election day, prevote, abroad voting)), which we will call panel aggregation approach hereafter.

To clarify how this approach works we reiterate the verbal discussion above with mathematical notations. For each constituency $c = 1, \cdots, C$, there are elements of polling stations $i(c) = 1(c), \cdots, I(c)$. Let $\mathcal{E}$, $\mathcal{P}$, and $\mathcal{AB}$ indicate the set of election day, prevote, and abroad polling stations. Let $N(c)$ and $N_{AB}(c)$ denote the total number of eligible voters except vot-

[10]We will briefly discuss Klimek, Yegorov, Hanel and Thurner (2012) in the discussions section.

[11]`https://www.yna.co.kr/view/GYH20200411001000044`

ers who preregistered for abroad voting, and voters who preregistered for abroad election in each constituency c. $V_{i(c)}$ stands for the number of valid votes for the polling station $i(c)$ of constituency c. Now we define the following objects as the number of valid votes of election day, prevote, and abroad voting. $\mathbb{1}$ is an indicator function.

$$E(c) = \sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{E}) V_{i(c)} \tag{2}$$

$$P(c) = \sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{P}) V_{i(c)} \tag{3}$$

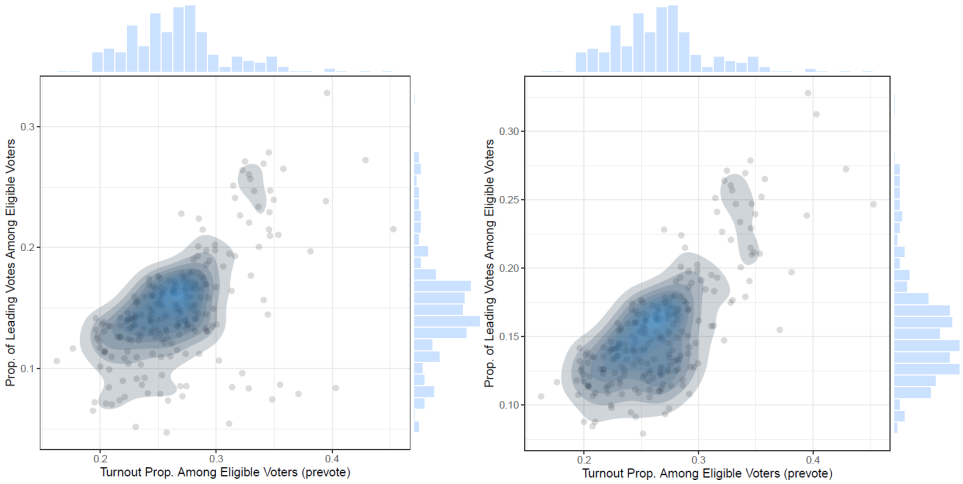
$$AB(c) = \sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{AB}) V_{i(c)} \tag{4}$$

Then, for each constituency c and voting method m=election day, prevote, and abroad election, the observed turnout proportion are expressed as

$$t_{cm} = \begin{cases} \frac{E(c)}{N(c)} & \text{if } m = \mathcal{E} \\ \frac{P(c)}{N(c)} & \text{if } m = \mathcal{P} \\ \frac{AB(c)}{N_{AB}(c)} & \text{if } m = \mathcal{AB} \end{cases} \tag{5}$$

Since election-day voting is held after prevoting, the "eligible voters" for election-day voting could also be expressed as $N(c) - P(c)$ instead of $N(c)$. Using $N(c)$ instead of $N(c) - P(c)$ is more consistent with the idea that prevoters and election day voters are samples drawn from an identical population, an assumption that is widely held by those who are suspicious of election frauds in the 2020 Korea Parliamentary Election. We will further address this issue in the replications section.

We aggregate mobile voting into election day voting. It is somewhat arbitrary and we might alternatively leave them as a separate unit of analysis. The reason why we aggregate them into election day voting units will become more clear in the later section where we discuss several advantages of panel aggregation approach. For reference, number of valid votes for mobile voting accounts for 0.32%(≈90,162/28,496,032) of total valid votes.



(a) Democratic Party specification (b) Constituency leader specification

Figure 2: Prevote data plots by panel aggregation approach

Notes: These figures are based upon panel aggregation approach suggested in this paper. Horizontal and vertical axis indicate the turnout proportion among eligible voters, and the vote rate of the (a)Democratic Party and (b)constituency winner candidates. The mean of prevote turnout rates across 253 constituencies is 26.68%.

The primary drawback of this approach is that aggregating granular data could obscure statistical anomalies that cannot be captured at the aggregate level. One could invoke ecological inference when we compare the results obtained from aggregation and those calculated by using polling station level data(Mebane, 2020b). Aggregating polling station-level data into a set of panel of constituencies and voting method might influence the low-key observation. Admittedly, there are at least three reasons why panel aggregation approach is more credible than polling station analysis in the context of the 2020 Korea election.

First, this panel aggregation approach enables us to make a consistent comparison of election-day voting and prevoting. Each voter is not assigned to a particular polling station for prevoting, while the voter can only vote at the designated polling station on election day. Considering this electoral process, the "eligible voters" for prevoting are the total number of eligible voters in each constituencies. Aggregating them into a panel of constituencies and voting method gives us one way of comparing election-day voting and prevoting in a coherent way.

Second, Mebane (2019b) mentions that the "observed aggregation unit"s i, "precinct" or

"polling station," are the units of analysis. In fact, Ferrari, McAlister and Mebane (2018) aggregates Brazilian polling station data and precinct data at the town level and controls for town-level covariates to predict τ_i. There is no clear theoretical condition under which aggregating polling station units are prohibited while i is denoted as "aggregation" units.

Third, the panel aggregation approach gives greater weights on the prevote polling stations in analysis. The biggest controversy of the Korean election is about prevote polling stations. Among 19,131 polling stations. 3,814 of them are prevote polling stations, which accounts for 19.94 % ($\approx$3,814/19,131) of the units of analysis, and Mebane (2020a) estimates that 25%-28.7% of prevote polling stations are fraudulently manipuated. By construction, on the other hand, prevote takes up 33%($\approx$1/3) of the units of analysis in the panel aggregation approach. If there was any large-scale election frauds in the prevote polling stations, this approach would be doing a good job of detecting frauds. According to Mebane (2020a)'s model, 1,234,217-1,418,079 votes out of 28,738,468 votes are fraudulently manipulated, and most of the frauds happened in the prevote polling stations. As long as the model is correctly specified, the panel aggregation approach would successfully detect election frauds in `eforensics` model. This point will become more clear when we look at the replication results under a set of different specifications.

Compared with You (2020), panel aggregation allows us to account for election frauds that could potentially be inherent in prevote units while he aggregates the polling stations data into a district(*eupmyeondong*) level. You (2020) also finds that the probability of no frauds in `eforensics` is reduced to 98.39% after calculating turnout rates at the district level.

Table 1 reports the replication results of prbability estimation under a vareity of specifications. Following Mebane (2020a), we set the first 5,000 draws as burn-in, the number of number of iterations in the adaptive phase as 1,000, the number of iterations in the sampling phase as 1,000, and the number of MCMC chains as 4.

Columns (1) and (6) are taken from Mebane (2020a). The probability that there is no fraud is around 92.7%-93.3%, and the probability that there is incremental fraud is 6.2%-6.5%.

Columns (2) and (7) document these probability estimation when we take panel aggre-

Table 1: Sensitivity analysis of Mebane (2020a)

	Democratic Party specification					Constituency leader specification				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Mebane method	✓		✓	✓		✓		✓	✓	
panel aggregation		✓	✓	✓	✓		✓	✓	✓	✓
no prevote dummy					✓					✓
mean prevote turnout	98.62%	26.68%	100%	98.76%	26.68%	98.62%	26.68%	100%	98.76%	26.68%
No fraud	0.933	0.987	0.817	0.840	0.987	0.927	0.982	0.815	0.829	0.986
Incremental fraud	0.062	0.012	0.174	0.152	0.011	0.065	0.017	0.174	0.165	0.013
Extreme fraud	0.005	0.001	0.009	0.008	0.001	0.008	0.001	0.011	0.006	0.001

Note: *Mebane method* indicates the approach where the turnout proportion of prevote unit are mistakenly set to be close to one or equal to one. Column (1) and (6) are results from Mebane (2020a). Column (2) and (7) are results from panel aggregation approach with correctly specified model. Column (3) and (8) are results from panel aggregation approach with 100% of prevote turnout proportion. Column (4) and (9) are estimated by aggregating the prevote turnout proportion without correcting for the misspecification in Mebane (2020a) so that $t_{cm} = [\sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{P}) V_{i(c)}] / [\sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{P})(V_{i(c)} + \delta_{i(c)})]$ for each constituency c. Column (5) and (10) are same as (2) and (7) except that we do not control for an indicator for prevote unit. Columns (1)-(5) are from Democratic Party specification and columns (6)-(10) are from constituency leader specification.

gation approach. The probability that there is no fraud is now 98.2%-98.7%. For context, Mebane (2019a) concludes that there is no fraud in the 2019 Bolivia Election and in the 2016 Korea Parliamentary Election with probability 99.09% and 97.9% of no fraud respectively.

Under panel aggregation approach, the intercept term is -0.833 and the fixed effect of prevote is -0.025. Then the estimated turnout proportion of prevote τ_i is around 29.78%($\approx \frac{1}{1+\exp(+0.833+0.025)}$), which is higher than the mean of observed prevote turnout proportion 26.68%.

These results of probability estimation could further change in either direction depending on whether we choose to control for constituency-level characteristics such as past election outcomes, income inequality, or the fraction of recipients of social welfare program instead of constituency fixed effect as implemented in Ferrari, McAlister and Mebane (2018) where they control for proportion of white, inequality, fraction of Blosa Familia receipients, and income per capita.

Columns (3) and (8) are estimated after intentionally setting the turnout proportion of prevote units as 100% from (2) and (7). Even after controlling for an indicator variable of prevote units, the probability of no fraud from `eforensics` model is 81.5%-81.7% and it is substantially lower than those of the panel aggregation approach with correct data.

Columns (4) and (9) exhibit results from panel aggregation while preserving the misspecification contained in Mebane (2020a). When aggregating the polling-station level data

into a panel of constituencies and voting method, the observed turnout rates of prevote units are calculated by $t_{cm} = \frac{\sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{P}) V_{i(c)}}{\sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{P})(V_{i(c)} + \delta_{i(c)})}$ for each constituency c. In this case, the mean of prevote turnout proportion is 98.76% and we control for an indicator for prevote units. The probability of no fraud is only about 82.9%-84%.

It is crucial to note that if the difference results from Mebane (2020a) and panel aggregation approach are purely driven by aggregation *per se*, then we would detect no election frauds in columns (3), (4), (8), and (9). Even after controlling for the fixed effect of prevote units, however, we detect significant amount of election frauds. If we do not fix faulty data, then panel aggregation approach still detects election frauds. This demonstrates that the difference between (1) and (5) versus (2) and (6) derives from misspecification in Mebane (2020a) rather than from aggregation that could potentially nullify low-key observations. The misspecification of Mebane (2020a) is that it uses prevote turnout rates of close to 100% as inputs. As long as this feature is maintained, election frauds are detected in `eforensics` model even after we take our aggregation approach. The narrative coming out of Mebane (2020a)'s statistical analysis is that ballot boxes in prevote polling stations seem to contain a large number of manufactured votes because the turnout rates of close to 100% are too high given that much of the other polling stations do not show turnout rates as high as 100%. This has nothing to do with the *observed* pattern in the data. If the turnout rates for prevoting is 100%, then the turnout rates for election day voting cannot be greater than 0%.

Apart from Mebane (2020a)'s analysis, one more subtle point that deserves attention is an idea that prevoters and election day voters are random samples that are independent and identically distributed, and therefore there should be no systematic difference in voting patterns between these two groups. This argument is made by those who are suspicious of election frauds in the 2020 Korea Election. This is a strong assumption that needs to be verified, but for now we suppose that this assumption holds[12]. If it is true, then it makes more sense not to control for an indicator for prevote, since the purpose of controlling for dummy variables is to account for qualitative difference between groups. If there is no difference between prevoters and election day voters, then an indicator for prevote

[12]Some evidence from Germany and Florida show that there are significant differences in voting patterns across voting methods(Wagner and Lichteblau, 2020; Herron and Smith, 2012)

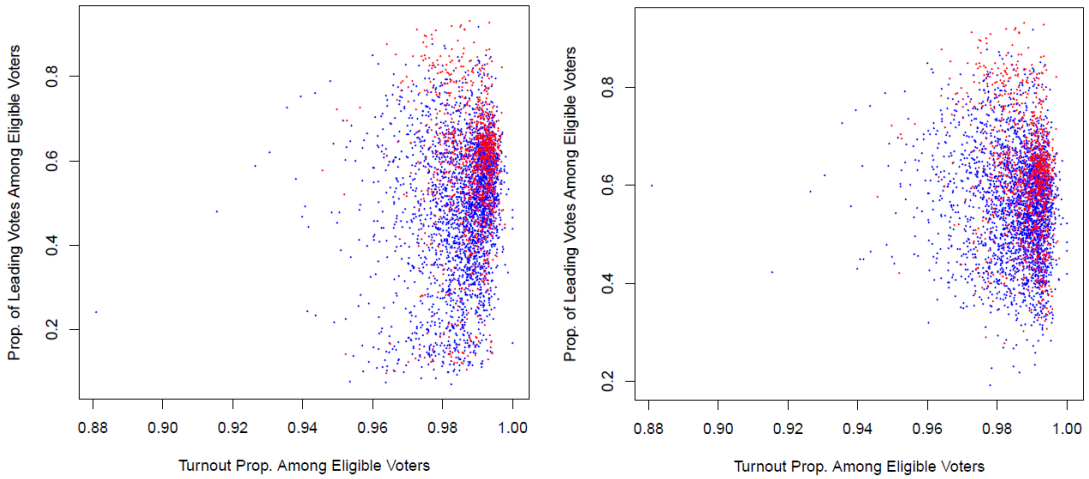
units is redundant at best. Columns (5) and (10) show that when we do not control for an indicator variable under panel aggregation approach, we also detect election frauds to a much smaller degree than Mebane (2020a).

Figure 3 depicts fraudulent and nonfraudulent units of anlaysis for prevote units in Mebane (2020a) and panel aggregation approach. The red dots among blue dots denote units of analysis that are suspected of being fraudulently manipulated. Unless misspecified, `eforensics` model does not detect large-scale election frauds in the data as much as Mebane (2020a) does. Figure 4 displays the same scatter plots of prevote units drawn by panel aggregation approach, but with misspecified models as implemented in Mebane (2020a). We can observe that prevote units are classified as "fraudulent" when the turnout rates are mistakenly set to be 100% or close to 100%. Figure 5 shows the entire units of analysis and that prevote units with turnout rates of 100% or close to 100% are classified as fraudulent units.

Comparing probability estimates from a different set of specifications illustrates three points. First, an indicator variable controlling for prevote does not fix statistical artifacts from misunderstanding of prevote turnout rates. Second, when there is an anomaly with prevote units, such as falsely setting prevote turnout rates as close to 100%, panel aggregation approach does a good job of detecting them when prevote units account for larger share of the total units of analysis. Third, aggregation alone does not nullify statistical anomalies, or "election frauds," that are detected because of confusions about the raw data at least in the context of the 2020 Korea Parliamentary Election.

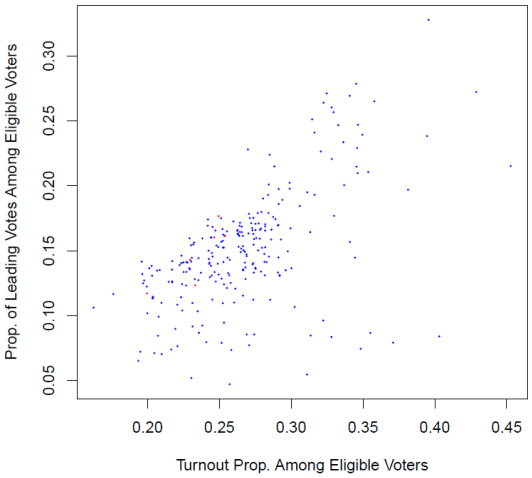
We do not document the number of manufactured votes or stolen votes. In the real world, only one fraudulent vote could challenge the legitimacy of elections. Yet, even a large number of "fraudulent votes" in `eforensics` model do not necessarily mean that election frauds occurred. A model is an approximation of the real world, and the meaning of "fraudulent votes" in the `eforensics` model can only be fully understood in each specific context. `eforensics` estimates that 10,723.3 votes and 26,452.5 votes are fraudulent in the first and the second round in the 2016 Austria election, but Mebane concludes that there was no election frauds because the numbers are less than the voting margin of 30,863. For Wisconsin in the 2016 US Presidential election, eforensics shows that 29,854.3 votes are

fraudulent, which he interprets as evidence for voter suppression rather than malfeasance. In Bolivia Election in 2019, `eforensics` estimates 22,519.8 votes are fraudulent but Mebane concludes that the election is not rigged since the fraudulent votes are less than the voting margin(Mebane, 2019a). In the 2008 U.S Presidential Election, 2,528 votes are estimated to be fraudulent in `eforensics` model, but again, this number is not large enough to turn the voting margin and he concludes that this pattern is driven "most likely" by "strategic activity," "while it is difficult to a develop a measure of strategic voting for California 2008 that can be associated with the fraud estimates(Mebane, 2016)."

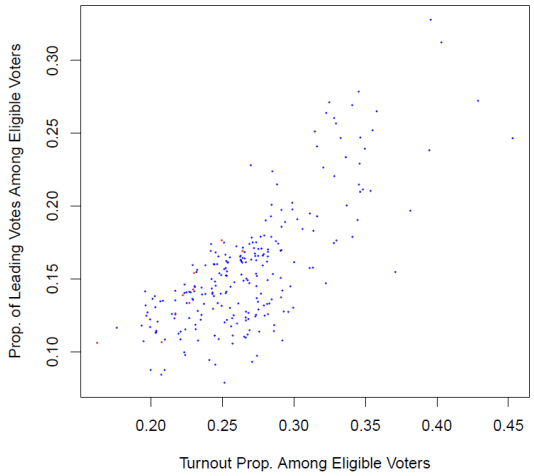


(a) prevote(Democratic Party)

(b) prevote(constituency leader)



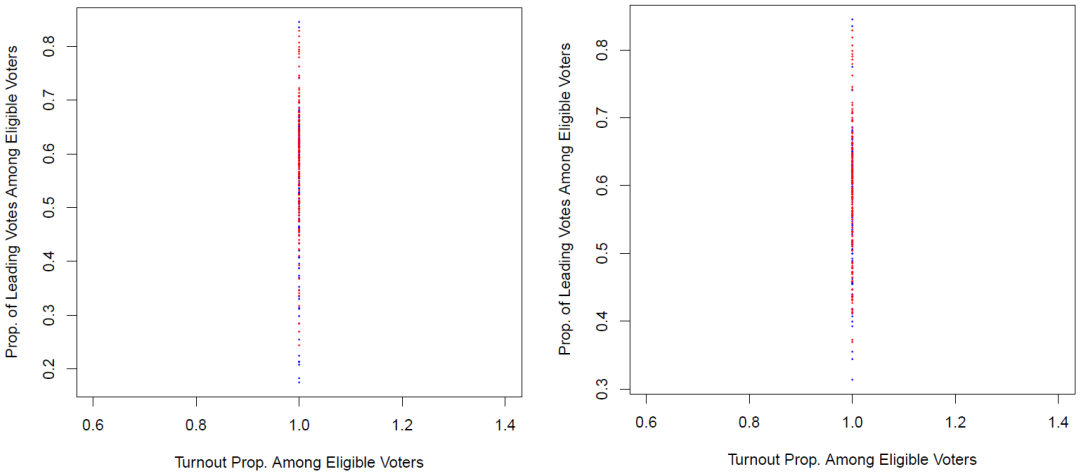
(c) prevote(Democratic Party)



(d) prevote(constituency leader)

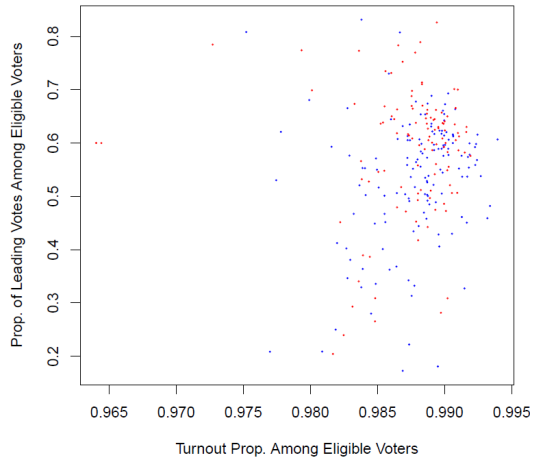
Figure 3: Prevote scatterplots: Main results

Notes: Blue dots and red dots denote nonfraudulent and fraudulent units of analysis. Figures (a) and (b) are taken from Mebane (2020a). Figures (c) and (d) are from panel aggregation approach with correctly specified models. Horizontal and vertical axis indicate the turnout proportion among eligible voters, and the vote rate of the Democratic Party((a), (c)) and constituency winner((b), (d)).

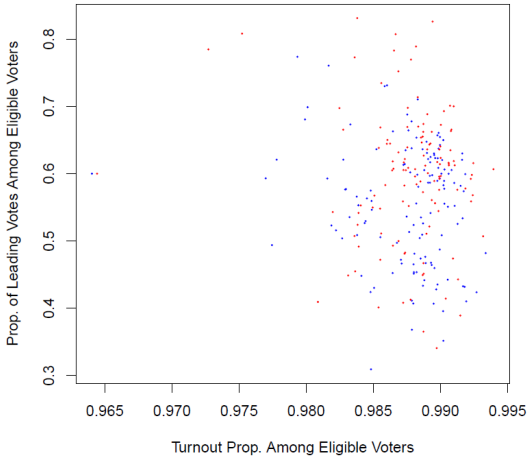


(a) prevote(Democratic Party)

(b) prevote(constituency leader)



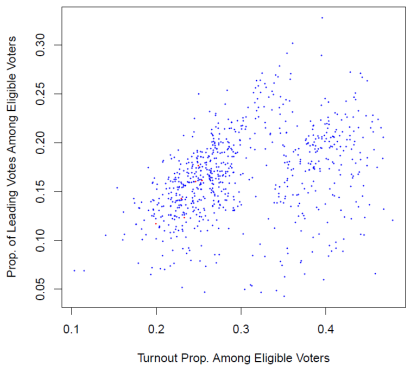
(c) prevote(Democratic Party)



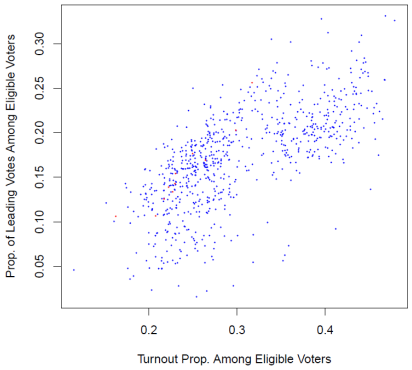
(d) prevote(constituency leader)

Figure 4: Prevote scatterplots: Panel aggregation with misspecification

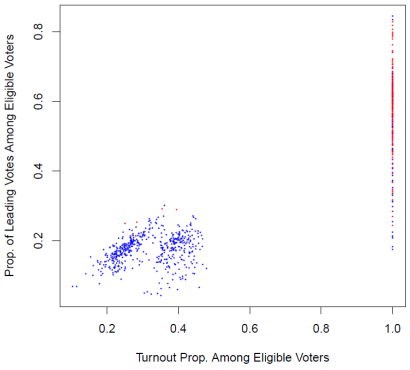
Notes: Blue dots and red dots denote nonfraudulent and fraudulent units of analysis. Figures (a) and (b) are taken from panel aggregation approach after replacing the number of eligible voters for prevote units as equal to the numebr of valid votes so that the prevote turnout rates are set to be 100%. Figures (c) and (d) are from panel aggregation approach after naively aggregating prevote units so that $t_{cm} = [\sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{P}) V_{i(c)}] / [\sum_{i(c)=1(c)}^{I(c)} \mathbb{1}(i(c) \in \mathcal{P})(V_{i(c)} + \delta_{i(c)})]$ for prevote units in the panel for each constituency c. Horizontal and vertical axis indicate the turnout proportion among eligible voters, and the vote rate of the Democratic Party((a), (c)) and constituency winner((b), (d)).



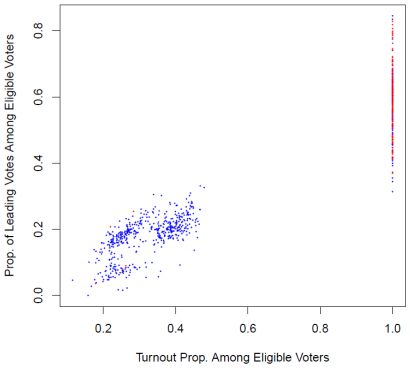
(a) Entire units(Democratic Party)



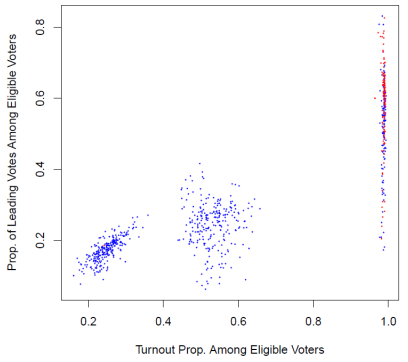
(b) Entire units(constituency leader)



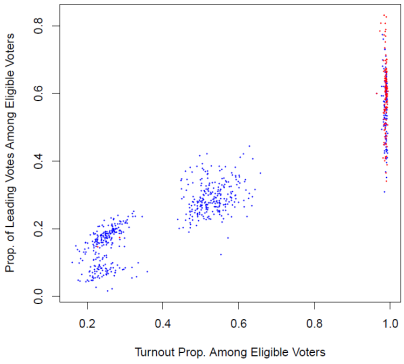
(c) Entire units(Democratic Party)



(d) Entire units(constituency leader)



(e) Entire units(Democratic Party)



(f) Entire units(constituency leader)

Figure 5: Entire units scatterplots

Notes: Blue dots and red dots denote nonfraudulent and fraudulent units of analysis. Figures (a) and (b) are taken from panel aggregation approach with corretly specified models. Figures (c) and (d) are from panel aggregation approach with prevote turnout rates of 100%. Figures (e) and (f) are from panel aggragation approach with naively aggregating prevote units. Horizontal and vertical axis indicate the turnout proportion among eligible voters, and the vote rate of the Democratic Party((a), (c), (e)) and constituency winner((b), (d), (f)).

3.2 Absentee/mail voting analysis

Mebane (2020b) suggests an alternative approach discussed in Ferrari and Mebane (2017). In short, the idea here is to account for the fact that voters who did not cast their votes on election day might have participated in the voting had they been given the opportunity to opt in the absentee voting and vice versa.

To express this idea with mathematical notations, suppose there are n_0 in-person polling stations and n_1 absentee polling stations. Each absentee polling stations correspond with one or more in-person polling stations. m_i is t he set of in-person polling stations associated with absentee polling stations $i = n_0 + 1, \cdots, n_0 + n_1$. The number of eligible voters for absentee polling station $i = n_0 + 1, \cdots, n_0 + n_1$ is $M_i = \sum_{k \in m_i} N_i$. φ_i is a parameter to be estimated that indicates the propensity to show up at the in-person polling stations instead of absentee polling stations. Then the observed turnout rates can be presented as

$$t_i = \begin{cases} \frac{V_i}{N_i \varphi_i} & \text{if } i = 1, \cdots, n_o \\ \frac{V_i}{\sum_{k \in m_i} N_k (1 - \varphi_k)} & \text{if } i = n_0 + 1, \cdots, n_0 + n_1 \end{cases} \tag{6}$$

Two issues needs to be resolved to implement this approach to analyze the Korean election. Again, understanding what prevote means and how they are presented in the raw data in Korea National Election Commission (2020) is a prerequisite for an sensible analysis. First, if we fail to understand the raw data in the first place, as we have discussed in the Section 2, we cannot correctly pin down N_i.

Second, according to Ferrari and Mebane (2017), "[e]ach in-person polling station is associated with at most one absentee station" while "some in-person stations may not be associated with an absentee station." For each constituencies there are several prevoting polling stations and contrary to most early voting in the US or advance poll in Canada, there is no geographical restrictions for prevoting in South Korea. Hence an in-person polling station is bound to be correspond with multiple prevote polling stations. We may get around this issue by aggregating polling-station data, an approach that is opposed by Mebane (2020b).

4 Discussions

Mebane (2019b)'s model is motivated by Klimek, Yegorov, Hanel and Thurner (2012). The red circles in Figure 6 display voting patterns that could have been driven by ballot box stuffing. This is similar to what Mebane (2020a) finds in the Korean data except that the "stuffed" units do not necessarily exhibit voting patterns in favor of the incumbent candidates nor constituency winners. Indeed, among 22 lawmakers that were elected owing to "election frauds" according to `eforensics` model, 7 of them are from the opposition party. However, as we have discussed, figures (a) and (b) in Figure 7 are not what the *real* observed data shows. This pattern exists only in misperceived data in Mebane (2020a). If we want to infer hidden behavior out of observed data, we need to have a clear understanding of what the observed data means.

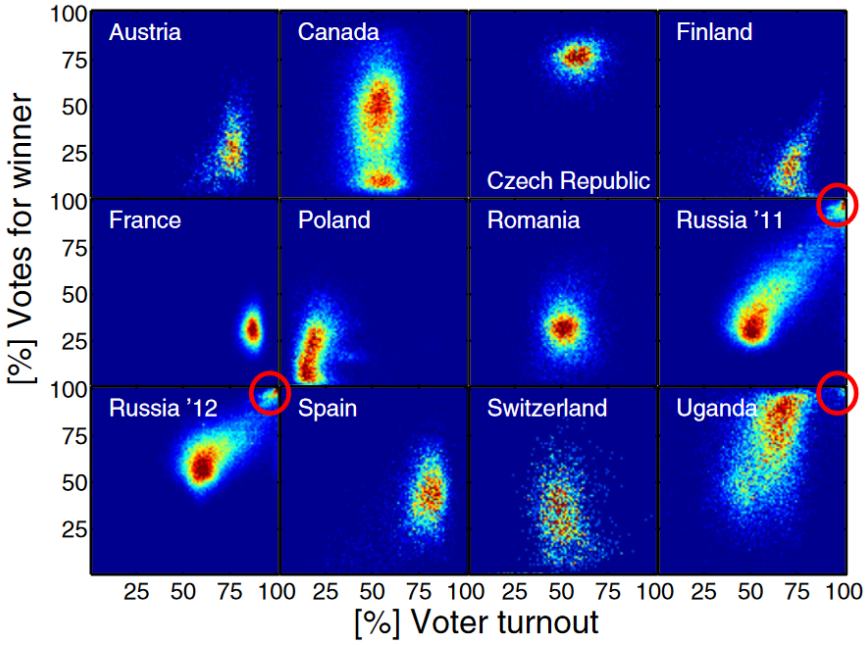
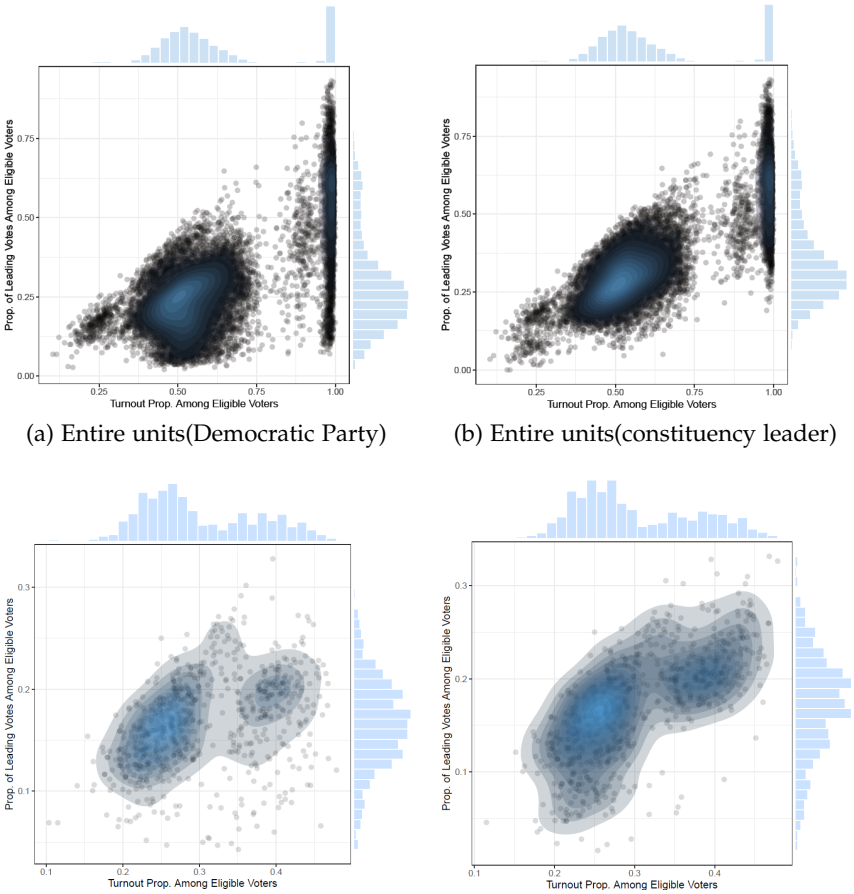


Figure 6: Election fingerprints

Notes: This figure is taken from Klimek, Yegorov, Hanel and Thurner (2012). Horizontal axis is voter turnout rates and vertical axis is a fraction of votes for a winning party or a candidate. Red circles indicate the aggregation units where election frauds might have occurred.

Figure 7 depicts total units of analysis that are drawn by Mebane (2020a) and panel aggregation approach with correctly specified models. After correcting for the confusion



(a) Entire units(Democratic Party)

(b) Entire units(constituency leader)

(c) Entire units(Democratic Party)

(d) Entire units(constituency leader)

Figure 7: Entire units scatterplots

Notes: Figures (a) and (b) are taken from Mebane (2020a). Figures (c) and (d) are from panel aggregation approach with correctly specified models. Horizontal and vertical axis indicate the turnout proportion among eligible voters, and the vote rate of the Democratic Party((a), (c)) and constituency winner((b), (d)).

revolving around prevote polling units, we do not the pattern of ballot box stuffing that is observed in figures (a) and (b) of Figure 7.

As Mebane (2020a) reminds us, there is a saying that all models are wrong but some are useful. We want to add that a model is not useful when we use inputs that are not required by the model. When we replace misspecified models with correctly specified ones, the probability of election frauds by `eforensics` model decreases substantially. Aggregating observed data could in principle conceal important information that can only be observed at the disaggregated level. Meanwhile, when we take our aggregation approach without

correcting for the misspecified models inherent in Mebane (2020a), we detect large-scale election frauds in `eforensics` model. "Election frauds" in Mebane (2020a) are not removed by aggregation itself when we intentionally keep misspecified models.

We do not take the replication result as evidence against election frauds. We do not discuss whether the model itself is doing a good job in distinguishing innocuous voting behavior from election frauds. Testing for the validity of `eforensics` model is beyond the scope of this paper although we may also evaluate Mebane (2019b)'s `eforensics` model in the broad context of forensic research in social sciences(Callen and Long, 2015; Enikolopov, Korovkin, Petrova, Sonin and Zakharov, 2013; Ritter, 2008; Zitzewitz, 2012).

References

Callen, Michael and James Long, "Institutional corruption and election fraud: Evidence from a field experiment in Afghanistan," *American Economic Review*, 2015, *105* (1), 354–81.

Deckert, Joseph, Mikhail Myagkov, and Peter Ordeshook, "Benford's Law and the detection of election fraud," *Political Analysis*, 2011, *19* (3), 245–268.

Enikolopov, Ruben, Vasily Korovkin, Maria Petrova, Konstantin Sonin, and Alexei Zakharov, "Field experiment estimate of electoral fraud in Russian parliamentary elections," *Proceedings of the National Academy of Sciences*, 2013, *110* (2), 448–452.

Ferrari, Diogo and Walter Mebane, "Developments in Positive Empirical Models of Election Frauds," 2017.

_ , Kevin McAlister, and Walter Mebane, "Developments in Positive Empirical Models of Election Frauds: Dimensions and Decisions," *2018 Annual Meeting of the Society for Political Methodology*, 2018.

Herron, Michael and Daniel Smith, "Souls to the polls: Early voting in Florida in the shadow of House Bill 1355," *Election Law Journal*, 2012, *11* (3), 331–347.

Klimek, Peter, Yuri Yegorov, Rudolf Hanel, and Stefan Thurner, "Statistical detection of systematic election irregularities," *Proceedings of the National Academy of Sciences*, 2012, *109* (41), 16469–16473.

Korea National Election Commission, "Eleciton Statistics Available at `http://info.nec.go.kr/`," 2020.

Mebane, Walter, "Comment on "Benford's Law and the detection of election fraud"," *Political Analysis*, 2011, *19* (3), 269–272.

_ , "Election Forensics: Frauds Tests and Observation-level Frauds Probabilities," 2016.

_ , "Evidence Against Fraudulent Votes Being Decisive in the Bolivia 2019 Election," 2019.

_ , "eforensics: A Bayesian Implementation of A Positive Empirical Model of Election Frauds," 2019.

_ , "Anomalies and Frauds in the Korea 2020 Parliamentary Election, SMD and PR Voting with Comparison to 2016 SMD," 2020.

_ , "Peronal email exchange," 2020.

New York State, "Governor Cuomo Signs Landmark Legislation Modernizing New York's Voting Laws. Available at https://www.governor.ny.gov/news/governor-cuomo-signs-landmark-legislation-modernizing-new-yorks-voting-laws," 2019.

Ritter, Jay, "Forensic finance," *Journal of Economic Perspectives*, 2008, *22* (3), 127–47.

Wagner, Aiko and Josephine Lichteblau, "Germany Going Postal? Comparing Postal and Election Day Voters in the 2017 German Federal Election," *German Politics*, 2020, pp. 1–24.

Wooldridge, Jeffrey, *Introductory econometrics: A modern approach*, Nelson Education, 2016.

Yonhap News, "4 · 15총선 지역별 사전투표율(최종)[Prevote turnout proportion by region(Final result)]. Available at https://www.yna.co.kr/view/GYH20200411001000044," 2020.

You, Kyungjoon, "A Study on the Fraud of the 21st National Assembly Election: Revisiting Mebane(2020)'s Working Paper. Available at https://blog.naver.com/gangnamforyou/221992767799," 2020.

Zitzewitz, Eric, "Forensic economics," *Journal of Economic Literature*, 2012, *50* (3), 731–69.