# Frauds in the Korea 2020 Parliamentary Election*

Walter R. Mebane, Jr.†

April 29, 2020

*Thanks to Hun Chung for highlighting the concerns with the election (as did several others) and for pointing to the dataset used in the analysis.

†Professor, Department of Political Science and Department of Statistics, Research Professor, Center for Political Studies, University of Michigan, Haven Hall, Ann Arbor, MI 48109-1045 (E-mail: wmebane@umich.edu).

The 2020 parliamentary election in Korea is controversial, with fraud allegations.

The statistical model implemented in `eforensics`[1] offers evidence that fraudulent votes occurred in the election that may have changed some election outcomes. The statistical model operationalizes the idea that "frauds" occur when one party gains votes by a combination of manufacturing votes from abstentions and stealing votes from opposing parties. The Bayesian specification[2] allows posterior means and credible intervals for counts of "fraudulent" votes to be determined both for the entire election and for observed individual aggregation units.

It is important to keep in mind that "frauds" according to the `eforensics` model may or may not be results of malfeasance and bad actions. How much estimated "frauds" may be produced by normal political activity, and in particular by strategic behavior, is an open question that is the focus of current research. Statistical findings such as are reported here should be followed up with additional information and further investigation into what happened. The statistical findings alone cannot stand as definitive evidence about what happened in an election.

Figure 1 shows the distribution of turnout and vote proportions across aggregation units.[3] Each turnout proportion is (Number Valid)/(Number Eligible), and each vote proportion is (Number Voting for Party)/(Number Eligible). The data include counts for $n = 19072$ units. 328 "abroad_office" observations have zero eligible voters but often a small number of votes—the largest number is 23—and are omitted from the plots. Figure 1(a) uses vote proportions defined based on Democratic Party votes, and Figure 1(b) uses vote proportions defined based on the votes received by the party with the most votes in each

[1] `https://github.com/UMeforensics/eforensics_public`

[2] Ferrari, McAlister and Mebane (2018) and `http://www.umich.edu/~wmebane/efslides.pdf`

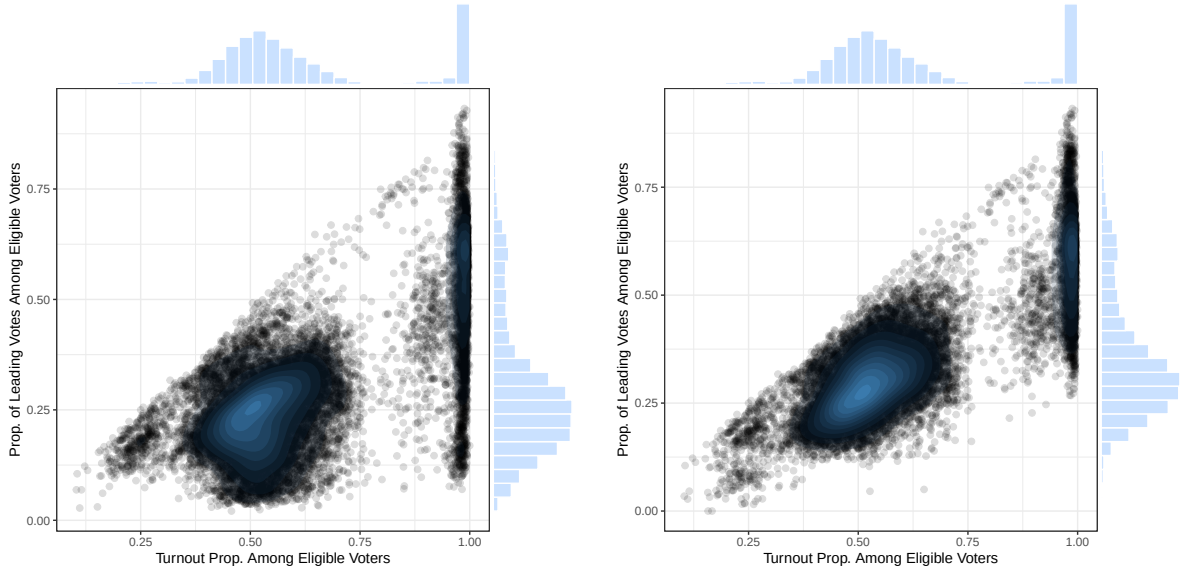
[3] Vote and eligible voter count data come from the file `korea_election_regional_21_eng.sqlite` at `https://gofile.io/?c=s0sqyW&fbclid=IwAR2w92Wq_QtcNxVn6K1HYlyEHnORV2yGYQGtCQU3oYf_OSSX7-tGARLSsSA`, from `https://gofile.io/?c=s0sqyW`, downloaded April 23, 2020 14:12. Constituency information is determined using the tables of "Electoral District and *Eupmyeon-dong*" at `http://info.nec.go.kr/main/showDocument.xhtml?electionId=0020200415&topMenuId=BI&secondMenuId=BIGI05` and the lists of winners at `http://info.nec.go.kr/main/showDocument.xhtml?electionId=0020200415&topMenuId=EP&secondMenuId=EPEI01`. Google Translate helped me by translating the Korean sources into English in my Chrome browser.

constituency. Fraud allegations have focused on the Democratic Party, but a principled way to analyze the single-member district election data is to consider that frauds potentially benefited the leading candidate in each constituency. In the figure differences between the two distributions are apparent, but both share a distinctive multimodal pattern. There appear to be clusters of observations that share distinctive levels of turnout and votes, some with low, medium, high and very high turnout. The diagonal edge feature in the plots results from using Number Eligible as the denominator for both proportions: when the party receives nearly all the valid votes, then the observation is near that diagonal.

Figure 1: Korea 2020 Parliamentary Election Data Plots

(a) Democratic Party

(b) Constituency leaders




Note: plots show turnout (number voting/number eligible) and vote proportions (number voting for party/number eligible) for (a) the Democratic Party or (b) the party the most votes in each constituency in aggregation units in the Korea 2020 parliamentary election. Plots show scatterplots with estimated bivariate densities overlaid, with histograms along the axes. 328 "abroad_office" observations reported with zero eligible voters but often with a positive number of votes are omitted.

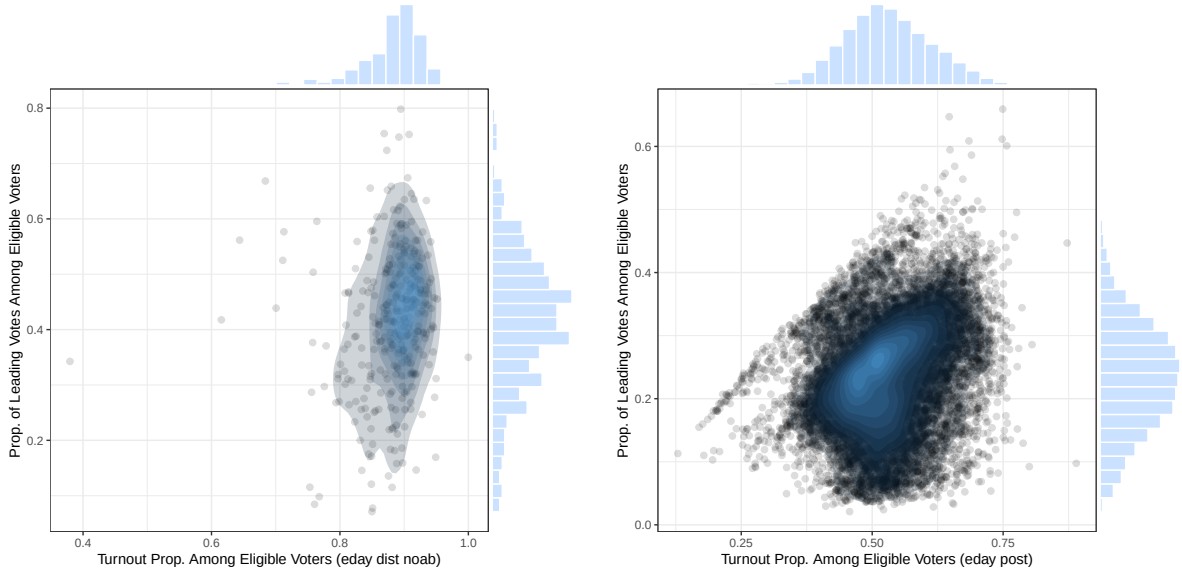
Figures 2 and 3 show that the different clusters in Figure 1 correspond with observations that are administratively distinctive. Figure 2 displays data for Democratic Party votes, and Figure 3 shows data for constituency leader votes. The four sets of units

that have distinctive distributions are district-level, election-day units that are not abroad (Figures 2(a) and 3(a)), postal, election-day units (Figures 2(b) and 3(b)), abroad units (Figures 2(c) and 3(c)) and pre-vote units (Figures 2(d) and 3(d)). Each subset of units (a), (b) and (d) has a mostly unimodal distribution: the marginal histograms are mostly near symmetric. But exceptional points are evident in each of these subsets. Abroad units are more distinctively bimodal when constituency leaders are considered than when the Democratic Party is in focus.

Figure 2: Korea 2020 Parliamentary Election Data Plots, Democratic Party

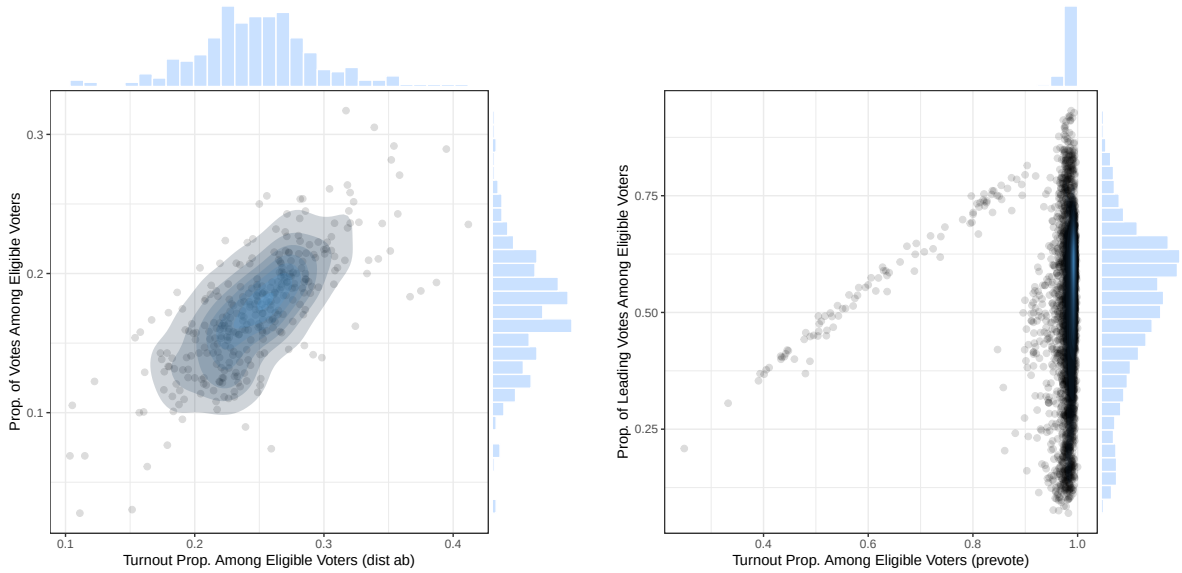
(a) district, election-day, not abroad

(b) postal, election-day




(c) abroad

(d) pre-vote

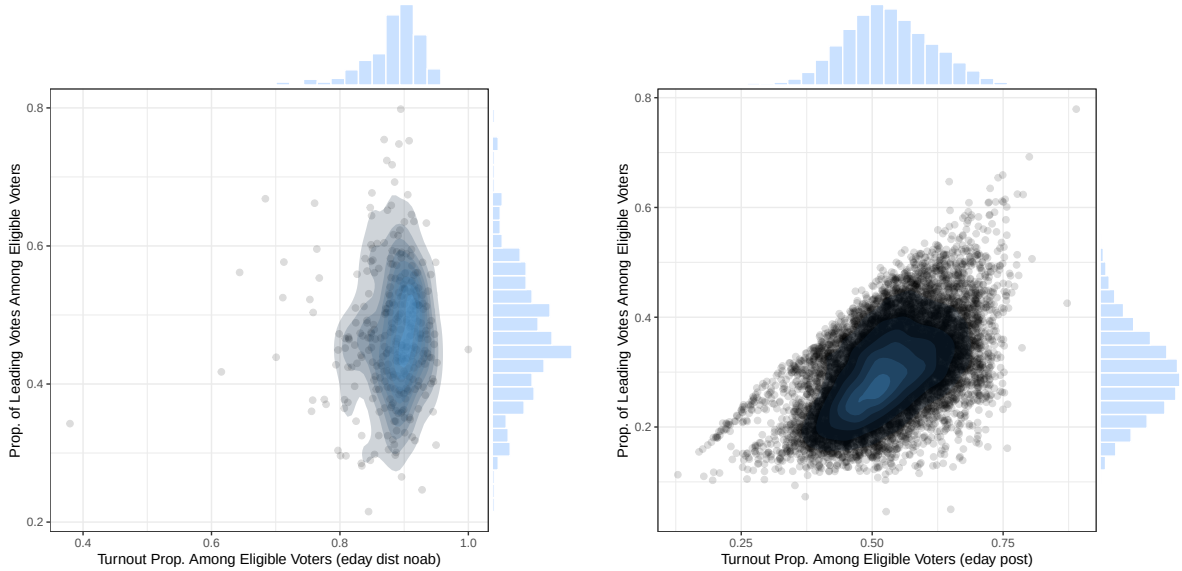



Note: plots show turnout (number voting/number eligible) and vote proportions (number voting for Democratic party/number eligible) for four subsets of observations: (a) district-level, election-day, not abroad; (b) postal election-day; (c) abroad; (d) pre-vote. Plots show scatterplots with estimated bivariate densities overlaid, with histograms along the axes. 328 "abroad_office" observations reported with zero eligible voters but often with a positive number of votes are omitted.

Figure 3: Korea 2020 Parliamentary Election Data Plots, Constituency Leaders

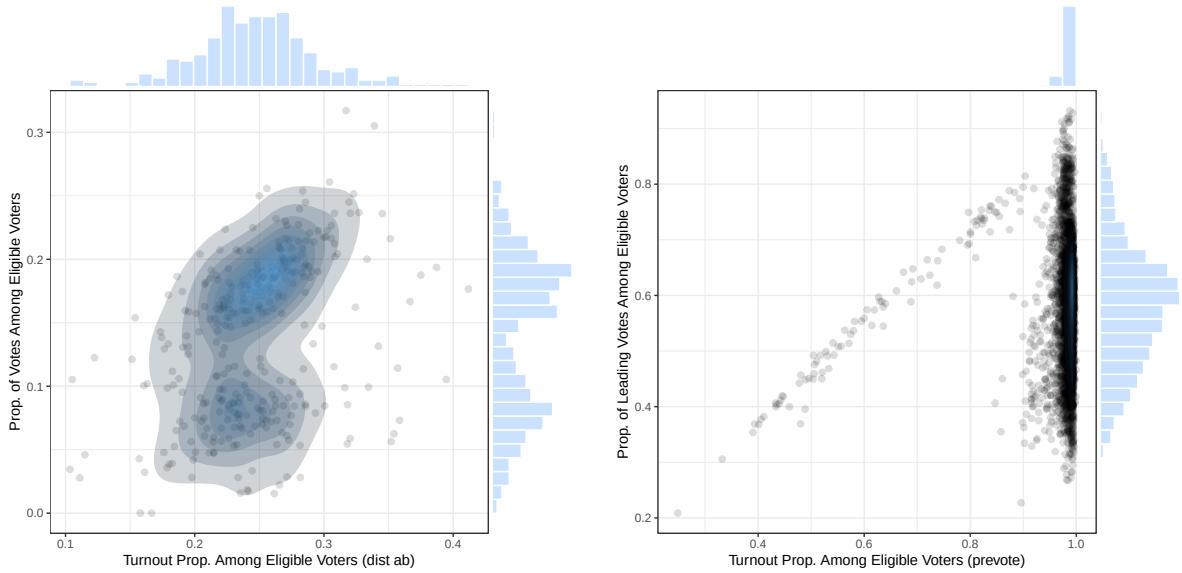
(a) district, election-day, not abroad

(b) postal, election-day




(c) abroad

(d) pre-vote




Note: plots show turnout (number voting/number eligible) and vote proportions (number voting for constituency-leading party/number eligible) for four subsets of observations: (a) district-level, election-day, not abroad; (b) postal election-day; (c) abroad; (d) pre-vote. Plots show scatterplots with estimated bivariate densities overlaid, with histograms along the axes. 328 "abroad_office" observations reported with zero eligible voters but often with a positive number of votes are omitted.

I estimate the `eforensics` model separately for the two definitions of leading party votes. Covariates for turnout and vote choice include indicators for pre-vote, postal, abroad and disabled-ship status and fixed effects for the 252 constituencies included in the data. The two specifications agree that 418 aggregation units are fraudulent, but 869 additional units are fraudulent in the Democratic party specification and 745 additional units are fraudulent in the constituency-leading party specification. As Table 1 shows, key parameter estimates are similar in the models. Parameters for the probabilities of frauds ($\pi_1$, $\pi_2$, $\pi_3$) are about the same between specifications, and coefficients for the turnout equation ($\tau_1$–$\tau_5$) are similar. Coefficients for vote choice ($\beta_1$–$\beta_4$) differ, reflecting the differences in vote proportions being modeled.

Figure 4 uses plots by subset of Democratic party focused observations to illustrate which observations are fraudulent according to the `eforensics` model with the Democratic party focused specification. Nonfraudulent observations are plotted in blue and fraudulent observations appear in red. The frequencies of fraudulent and not fraudulent units appear in the note at the bottom of the figure. Visually and by the numbers, frauds occur most frequently for pre-vote units (43.1% are fraudulent), next most frequently for for district-level, election-day, not abroad unts (3.14% fraudulent) then next most frequently postal election day units (.925% are fraudulent). None of the abroad units are fraudulent.

Figure 5 uses plots by subset of constituency-leader focused observations to illustrate which observations are fraudulent according to the `eforensics` model with the constituency-leader focused specification. Nonfraudulent observations are plotted in blue and fraudulent observations appear in red. The frequencies of fraudulent and not fraudulent units appear in the note at the bottom of the figure. Visually and by the numbers, frauds occur most frequently for pre-vote units (22.6% are fraudulent), next most frequently for postal election day units (2.09% are fraudulent) then next most frequently for district-level, election-day, not abroad unts (.920% fraudulent). None of the abroad units are fraudulent.

Table 1: Korea 2020 Parliamentary `eforensics` Estimates

(a) Democratic Party specification

| Parm. | Covariate | Mean | HPD.lo[a] | HPD.up[b] |
|---|---|---|---|---|
| $\pi_1$ | No Fraud | .928 | .924 | .931 |
| $\pi_2$ | Incremental Fraud | .0661 | .0624 | .0696 |
| $\pi_3$ | Extreme Fraud | .00588 | .00478 | .00690 |
| $\gamma_1$ | (Intercept) | .738 | .712 | .765 |
| $\gamma_2$ | pre-vote | 1.02 | .957 | 1.10 |
| $\gamma_3$ | postal | −.0347 | −.0409 | −.0269 |
| $\gamma_4$ | abroad | −.0365 | −.0411 | −.0310 |
| $\gamma_5$ | disabled-ship | .0475 | .0419 | .0539 |
| $\beta_1$ | (Intercept) | −.116 | −.137 | −.0944 |
| $\beta_2$ | pre-vote | .0473 | .0412 | .0560 |
| $\beta_3$ | postal | −.130 | −.149 | −.114 |
| $\beta_4$ | abroad | .203 | .190 | .214 |
| $\beta_5$ | disabled-ship | −.0513 | −.0597 | −.0388 |

(b) constituency leader specification

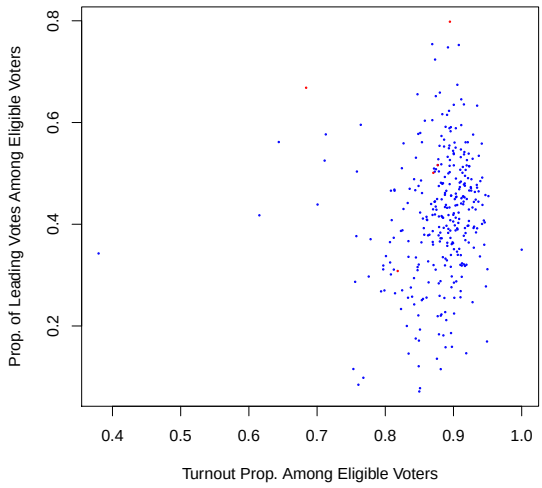
| Parm. | Covariate | Mean | HPD.lo[a] | HPD.up[b] |
|---|---|---|---|---|
| $\pi_1$ | No Fraud | .929 | .924 | .933 |
| $\pi_2$ | Incremental Fraud | .0648 | .0595 | .0697 |
| $\pi_3$ | Extreme Fraud | .00667 | .00553 | .00782 |
| $\gamma_1$ | (Intercept) | .692 | .671 | .714 |
| $\gamma_2$ | pre-vote | 1.12 | 1.05 | 1.19 |
| $\gamma_3$ | postal | −.0322 | −.0349 | −.0295 |
| $\gamma_4$ | abroad | −.106 | −.109 | −.103 |
| $\gamma_5$ | disabled-ship | .0467 | .0388 | .0582 |
| $\beta_1$ | (Intercept) | .202 | .193 | .211 |
| $\beta_2$ | pre-vote | −.0568 | −.0776 | −.0429 |
| $\beta_3$ | postal | .0696 | .0647 | .0745 |
| $\beta_4$ | abroad | −.00791 | −.0140 | −.00218 |
| $\beta_5$ | disabled-ship | −.00815 | −.0132 | −.00310 |

Note: selected `eforensics` model parameter estimates. Constituency fixed effects are not shown. For parameter notation see `http://www.umich.edu/~wmebane/efslides.pdf`. $n = 18744$.

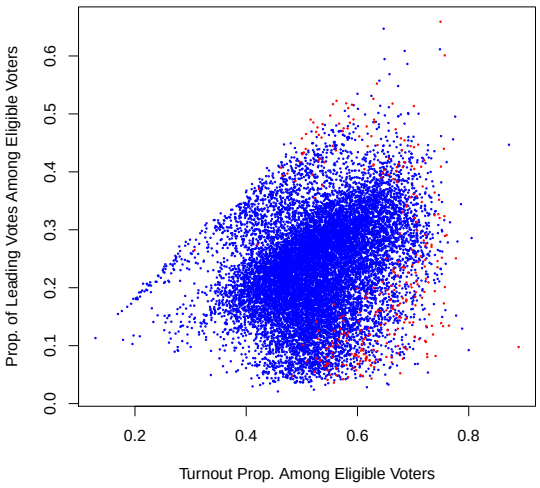
[a] 95% highest posterior density credible interval lower bound. [b] 95% highest posterior density credible interval upper bound.

Figure 4: Korea 2020 Fraud Plots , Democratic Party

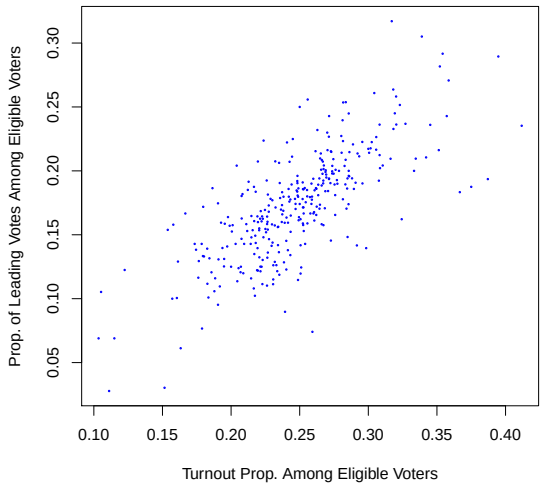
(a) district, election-day, not abroad



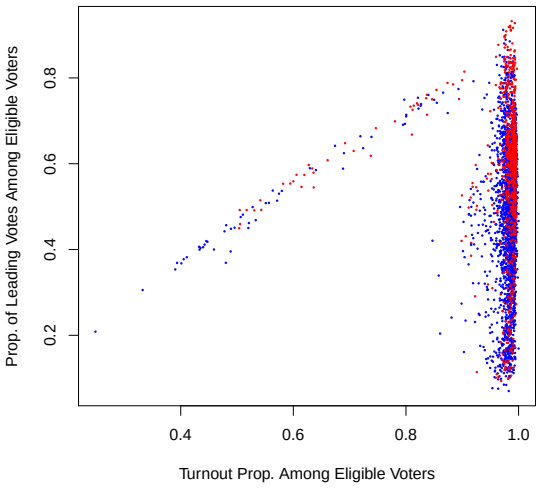

(b) postal, election-day




(c) abroad



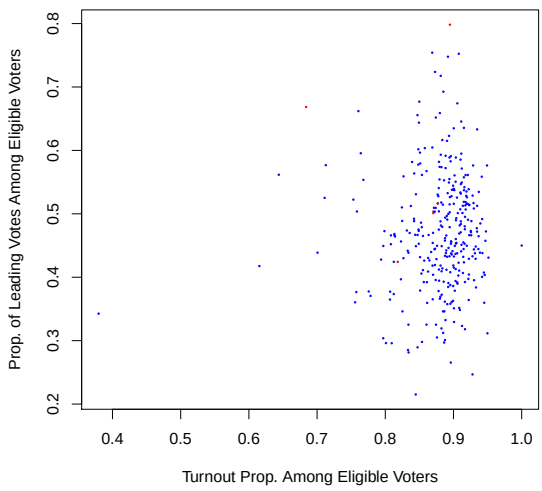

(d) pre-vote



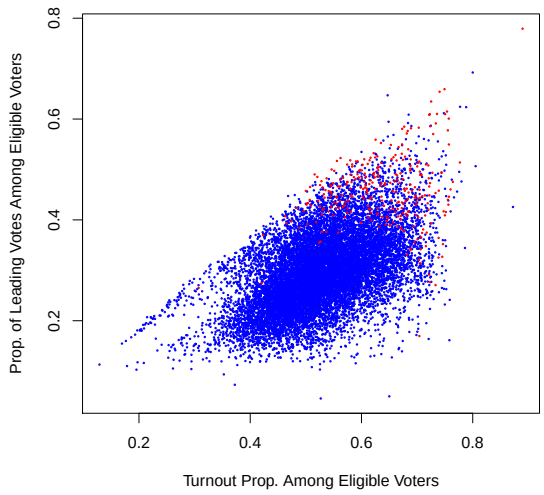

Note: plots show turnout (number voting/number eligible) and vote proportions (number voting for Democratic Party/number eligible) for four subsets of observations: (a) district-level, election-day, not abroad (10 fraudulent, 318 not); (b) postal election-day (131 fraudulent, 14155 not); (c) abroad (0 fraudulent, 328 not); (d) pre-vote (1146 fraudulent, 2656 not). Plots show scatterplots with nonfraudulent observations in blue and fraudulent observations in red. 328 "abroad_office" observations reported with zero eligible voters but often with a positive number of votes are omitted.

Figure 5: Korea 2020 Fraud Plots , Constituency Leaders

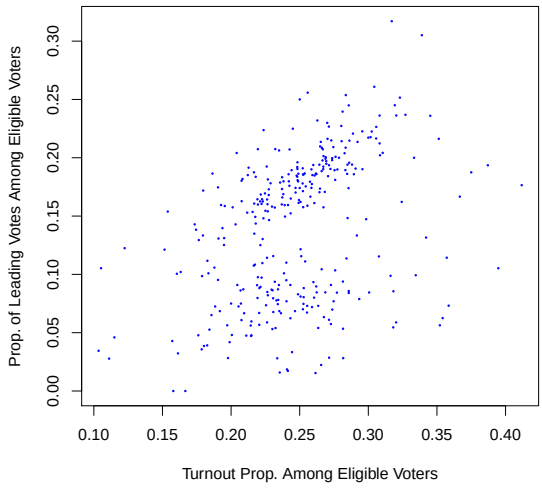
(a) district, election-day, not abroad



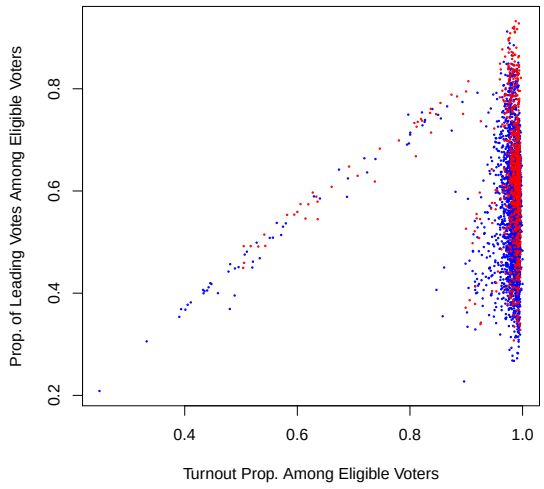

(b) postal, election-day




(c) abroad




(d) pre-vote




Note: plots show turnout (number voting/number eligible) and vote proportions (number voting for constituency-leading party/number eligible) for four subsets of observations: (a) district-level, election-day, not abroad (5 fraudulent, 323 not); (b) postal election-day (298 fraudulent, 13988 not); (c) abroad (0 fraudulent, 328 not); (d) pre-vote (860 fraudulent, 2942 not). Plots show scatterplots with nonfraudulent observations in blue and fraudulent observations in red. 328 "abroad_office" observations reported with zero eligible voters but often with a positive number of votes are omitted.

I use a counterfactual method to calculate how many votes are fraudulent.[4] Table 2 reports the observed counts of eligible voters, valid votes and votes for the (a) Democratic party and (b) constituency-leading party totaled over all units in the analysis, along with fraudulent vote count totals. The total of "manufactured" votes is reported separately from the total number of fraudulent votes: manufactured votes are votes that the model estimates should have been abstentions but instead were observed as votes for the leading party. Both posterior means and 95% and 99.5% credible intervals are reported. The results show that for the Democratic Party focused specification over all about 1,491,548 votes are fraudulent, and of the fraudulent votes about 1,122,169 are manufactured (the remaining 369379 are stolen—counted for the leading party when they should have been counted for a different party). Overall, according to the `eforensics` model, about 10.43% of the votes for the Democratic Party candidates are fraudulent. The results show that for the constituency-leading focused specification over all about 1,171,734 votes are fraudulent, and of the fraudulent votes about 910,444 are manufactured (the remaining 261,290 are stolen—counted for the leading party when they should have been counted for a different party). Overall, according to the `eforensics` model, about 7.26% of the votes for the constituency-leading candidates are fraudulent.

Fraudulent vote occurrence varies over constituencies. Counts of frauds by aggregation unit appear in a supplemental file[5], but I use the unit-specific fraudulent vote counts from the constituency-leader focused specification to assess whether the number of fraudulent votes is ever large enough apparently to change the winner of a constituency contest. For 236 constituencies it is not, but for 16 constituencies the number of fraudulent votes is large enough apparently to change the winner of the constituency contest. In 9 instances the apparently fraudulently winning party is the Democratic Party, in 6 instances it is the

[4]For a description of the method see "approach two" described at `http://www.umich.edu/~wmebane/efslides.pdf`.

[5]See the original **R** output files `wrkef2a_Korea2020AC_1d.Rout` and `wrkef2a_Korea2020aAC_1d.Rout` in `Korea2020ef.zip` for the numbers of fraudulent votes at each aggregation unit.

Table 2: Korea 2020 `eforensics` Estimated Fraudulent Vote Counts

(a) Democratic Party specification fraudulent counts

Observed Counts

| Voters | Valid | Votes |
|---|---|---|
| 43794881 | 28494664 | 14297282 |

| | 95% interval | | 99.5% interval | |
|---|---|---|---|---|
| Manufactured | lo | up | lo | up |
| 1122169.4 | 1085696.8 | 1162389.8 | 605047.5 | 1181520.5 |

| | 95% interval | | 99.5% interval | |
|---|---|---|---|---|
| Total | lo | up | lo | up |
| 1491547.9 | 1456551.0 | 1529447.6 | 1130549.7 | 1543719.3 |

(b) constituency leader specification fraudulent counts

Observed Counts

| Voters | Valid | Votes |
|---|---|---|
| 43794881 | 28494664 | 16144759 |

| | 95% interval | | 99.5% interval | |
|---|---|---|---|---|
| Manufactured | lo | up | lo | up |
| 910443.8 | 866426.2 | 950106.5 | 466261.7 | 964253.0 |

| | 95% interval | | 99.5% interval | |
|---|---|---|---|---|
| Total | lo | up | lo | up |
| 1171734.5 | 1117076.5 | 1211617.4 | 875150.9 | 1225551.3 |

Note: observed counts and total fraud posterior means and credible intervals based on `eforensics` model estimates.

United Future Party and in the remaining instance it is an Independent candidate.[6]

Given two specifications, which one is better? Probably neither model is correct, strictly speaking, even beyond the generality that no model is ever correct, but some are useful. If frauds only ever benefit the Democratic Party, then those frauds may have

[6]The particular constituencies that have these conditions can be identified by matching constituencies sequentially using the alphabet in "list of winners" tables available from `http://info.nec.go.kr/main/showDocument.xhtml?electionId=0020200415&topMenuId=EP&secondMenuId=EPEI01` (as of April 27, 2020 18:04 EST): Gangwon-do E (4367.5 fraudulent), Gyeonggi-do H (6622.1 fraudulent), Gyeonggi-do I (6629.7 fraudulent), Gyeonggi-do JJ (8512.9 fraudulent), Gyeonggi-do RR (7628.9 fraudulent), Gyeongsangnam-do E (2479.4 fraudulent), Daejeon B (4345.7 fraudulent), Daejeon G (4211.5 fraudulent), Busan G (3134.7 fraudulent), Busan H (3339.9 fraudulent), Seoul D (7727.9 fraudulent), Seoul F (6762.0 fraudulent), Seoul SS (3959.5 fraudulent), Incheon Metropolitan City A (4916.2 fraudulent), Incheon Metropolitan City D (2920.8 fraudulent), Chungcheongnam-do F (1809.9 fraudulent).

induced apparent frauds when we constrain frauds to benefit only constituency-leading candidates, because many of these do not affiliate with the Democratic Party. Similarly if only constituency-leading candidates benefit from frauds, then `eforensics` may be producing misleading results when we constrain frauds to benefit only the Democratic Party. Or perhaps other candidates—or several in each constituency—benefit from frauds and both specifications are producing misleading results. Possibly, of course, there are no frauds and something else is going on.

Caveats are many. The most basic caution is to keep in mind that "frauds" according to the `eforensics` model may or may not be results of malfeasance and bad actions. If some normal political situation makes the apparently fraudulent aggregation units appear fraudulent to the `eforensics` model and estimation procedure, then the frauds estimates may be signaling that "frauds" occur where in fact something else is happening. In particular there maybe something benign that leads many of the pre-vote units to have a turnout and vote choice distribution that differs so much especially from the distribution for election-day postal units, the latter comprising the bulk of the data. Likewise something benign may distinguish the election-day postal units that the `eforensics` model identifies as fraudulent. Beyond that general caution, there may something about the particular data used for the analysis that triggers the "fraud" findings—for instance, the data appear to be missing about 100,000 votes and one entire constituency, and the vote totals in the data for constituency-leading candidates do not always match totals reported in "lists of winners." And there may be something about the model specification that should be improved that would produce different results.

Statistical findings such as are reported here should be followed up with additional information and further investigation into what happened. The statistical findings alone cannot stand as definitive evidence about what happened in the election.

# References

Ferrari, Diogo, Kevin McAlister and Walter R. Mebane, Jr. 2018. "Developments in Positive Empirical Models of Election Frauds: Dimensions and Decisions." Presented at the 2018 Summer Meeting of the Political Methodology Society, Provo, UT, July 16–18.