Election Forensics Study of 2019 Democratic Republic of Congo Election

Walter R. Mebane, Jr.*

January 25, 2019

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

1 Introduction

I describe election forensics analysis (see Hicken and Mebane 2015) of polling site data leaked from “the Independent National Electoral Commission (CENI), representing 86% of the total votes cast” (Stearns 2019). The leaked data contain observations for 17,782 polling sites. Variables included in the data are as follows: province_id, province, clcr_id, clcr, circonscription_id, circonscription, siege, site_vote_id, nom_sv, adresse_sv, bv_prevus, bv_traites, electeurs_attendus, votants, candidat_id, voix. electeurs_attendus reports the number of eligible/registered voters and votants reports the number of voters participating at each polling site. voix reports the number of votes at each polling site for the candidate named in the variable candidat_id. The values of candidat_id that are associated with nonzero vote counts are X1001_44_10, X1001_48_14 and X1001_84_158. I don’t know which actual candidates correspond to the candidate ID codes.

Missing data: Data are missing for electeurs_attendus and votants for two polling sites, and vote counts are missing for 2,095 polling sites (including the two sites that have missing values for electeurs_attendus and votants). These observations with missing data are omitted from the analysis reported here. After omitting these observations, 15,687 observations remain.

The totals of electeurs_attendus, votants and the counts for the three candidates are

electeurs_attendus	39,766,849
votants	15,739,388
1001_44_10	2,974,455
1001_48_14	9,321,680
1001_84_158	2,909,489

Evidently candidate X1001_48_14 is reported as having vastly more votes at the nonmissing, leaked polling sites than are the other two candidates.

Election Forensics Analysis: A caveat on all the following analysis is that I know nothing about why data were leaked for these polling sites and not others, nor why data are missing for some polling sites and not others. Stearns (2019) gives reasons to believe the data "reflect the real results of the elections," but what he says is all I know about that. Even if the particular polling site counts do match the values for those sites in the official data, it is possible that the selection of leaked and nonmissing sites affects the analysis. Especially because the imbalance among candidates is much greater than was reported by the electoral commission,[1] the results from analyzing the complete collection of polling site data would likely differ in unknown ways.

Table 1 reports basic diagnostic results from the Election Forensics Toolkit (EFT) (Hicken and Mebane 2015; Mebane 2015). See Hicken and Mebane (2015) for descriptions of the particular test statistics. Hicken and Mebane (2015) suggest that having many of the statistics differ significantly from their nominal values is evidence that election frauds have occurred. In Table 1 almost every statistic for the vote counts differs from nominal values (the significantly deviating values appear highlighted in red), and one statistic deviates from nominal values for turnout. Notably the "`P05s`" statistic that shows a significant departure for turnout is a statistic that is hard to imagine occurring by chance. Rather it arises when agents who are manipulating votes want to be detected doing so (Kalinin 2017) or due to several agents who manipulate votes but imperfectly coordinate their actions (Rundlett and Svolik 2016).

No single indicator is perfect as a marker for election frauds, but the `P05s` measure is one of the least ambiguous indicators. The suggestion that votes are manipulated is reinforced by the results of the "spikes" test, which assesses whether patterns in the digits of vote proportions really are unusual or whether any pattern is merely a consequence of unmanipulated random counts (Rozenas 2017). For this analysis one must select one

[1] "The results contradict those published by the election commission on January 10th 2018, which proclaimed Felix Tshisekedi the winner with 38,57% of the vote, followed by Martin Fayulu with 34,8% and Emmanuel Ramazani Shadary with 23,8%" (Stearns 2019).

candidate as the candidate that may be benefiting by having votes fraudulently added to the candidate's vote totals at some polling sites (or otherwise favorably manipulated). I use the candidate with most votes in the leaked data, candidate `1001_48_14`, for this candidate that may have benefitted from frauds. The method assesses patterns in the proportion of votes counted for this candidate, using the turnout proportions to develop a statistical baseline. See Rozenas (2017) for details. For a polling site to be included in the analysis, we require that the number of eligible/registered votes be at least as large as the number participating, and that the number of voters participating be at least as large as the total number of votes across all candidates. Of the 15687 polling sites with no missing count data, 369 have more voters participating than there are eligible/registered voters, and one polling site has more votes cast for candidates than there are voters participating. Omitting these observations leaves $n = 15317$ polling sites to be analyzed using the "spikes" test.

As Figure 1 shows, the "spikes" test shows there is an excess of vote proportions for candidate `1001_48_14` that end in zero. Such a finding directly reinforces the results of the `P05s` test. The kinds of manipulations the `P05s` and "spikes" tests suggest occurred also indirectly explain why the other EFT tests give results that deviate significantly from nominal values: probably all the candidates' vote counts have been manipulated.

Last I have estimates from the likelihood finite mixture model (Mebane 2016). Here again `1001_48_14` is treated as the candidate that may have benefitted from frauds, and the number of polling sites in the analysis is $n = 15,317$. Table 2 shows that the likelihood ratio test statistic for the hypothesis that there are no frauds is 14855.2: comparing that to the chi square distribution with four degrees of freedom implies the hypothesis is rejected. Incremental fraud is estimated to occur with an unconditional probability of $\hat{f}_{\mathrm{i}} = .120$, and extreme fraud is estimated to occur with an unconditional probability of $\hat{f}_{\mathrm{e}} = .000131$. So, roughly speaking, incremental fraud is estimated to occur in twelve percent of polling sites. Because the parameter α is estimated to be greater than 1.0, i.e. $\hat{\alpha} = 1.40$, the frauds are estimated to involve more vote manufacturing than vote stealing. According to the model, the expected number of fraudulent votes counted for candidate 1001_48_14 is 512,454 votes from incremental fraud and 1,140 votes from extreme fraud. These counts of fraudulent votes are respectively 3.35 and .00746 percent of the reported votes (votants) at the included polling sites, and they are 4.09 and .00910 percent of the reported votes cast for the candidates.

Conclusion: Analysis of the leaked polling site data conveys an impression of an election replete with large frauds. Turnout figures and vote counts appear to have been manipulated in ways that are readily identifiable: patterns in the digits suggest that those doing the manipulations may want to get credit for having done them. The number of fraudulent votes is estimated to be in the hundreds of thousands—this number depends on the particular model used to determine it, but it is reasonable to believe at least that the frauds involve many purported votes.

Caveats are many. As previously mentioned, the fact that we have data from a selection of leaked and nonmissing polling sites means that we do not know whether the various estimates reflect the patterns that occur throughout the entirety of the data. Perhaps all the suspicious results are due to having a special selection of polling sites. Other limitations include that the analysis assumes all polling sites are effectively homogeneous. But in most settings polling sites are not homogeneous, and omitted factors such as consequential covariates that are not included in the analysis may be a reason for the results (cf. e.g. Mebane 2017). While the likelihood finite mixture model cannot effectively condition on such covariate information, related models currently nearing availability do (Ferrari, McAlister and Mebane 2018). Finally it is possible, although perhaps not most likely, that benign strategic behavior by voters is causing some of the methods mistakenly to signal that there is fraudulent behavior: but neither P05s nor the "spikes" test can be explained that way, and the finite mixture model results probably cannot be explained that way.

For a full forensic analysis of the election it would be good to have complete polling site data, along with the detailed data from the parallel vote tabulation that was conducted (Stearns 2019). Having observations for covariates that are believed to be associated with the diversity of relevant preferences or strategies would also be good.

References

Ferrari, Diogo, Kevin McAlister and Walter R. Mebane, Jr. 2018. “Developments in Positive Empirical Models of Election Frauds: Varying Dimensions.” Paper presented at the 2018 Summer Meeting of the Political Methodology Society, Provo, UT, July 19–21, 2018.

Hicken, Allen and Walter R. Mebane, Jr. 2015. “A Guide to Election Forensics.” Working paper for IIE/USAID subaward #DFG-10-APS-UM, “Development of an Election Forensics Toolkit: Using Subnational Data to Detect Anomalies.” URL: `http://www-personal.umich.edu/~wmebane/USAID15/guide.pdf`.

Kalinin, Kirill. 2017. “The Essays on Election Fraud in Authoritarian Regimes.” PhD thesis, University of Michigan.

Mebane, Jr., Walter R. 2015. “Election Forensics Toolkit DRG Center Working Paper.” Working paper for IIE/USAID subaward #DFG-10-APS-UM, “Development of an Election Forensics Toolkit: Using Subnational Data to Detect Anomalies.” URL: `http://www-personal.umich.edu/~wmebane/USAID15/report.pdf`.

Mebane, Jr., Walter R. 2016. “Election Forensics: Frauds Tests and Observation-level Frauds Probabilities.” Paper presented at the 2016 Annual Meeting of the Midwest Political Science Association, Chicago, April 7–10, 2016.

Mebane, Jr., Walter R. 2017. “Anomalies and Frauds(?) in the Kenya 2017 Presidential Election.” Working paper, `http://www-personal.umich.edu/~wmebane/Kenya2017.pdf`.

Rozenas, Arturas. 2017. “Detecting Election Fraud from Irregularities in Vote-Share Distributions.” *Political Analysis* 25(1):41–56.

Rundlett, Ashlea and Milan W. Svolik. 2016. “Deliver the Vote! Micromotives and Macrobehavior in Electoral Fraud.” *American Political Science Review* 110(1):180–197.

Stearns, Jason. 2019. “Who Really Won the Congolese Elections?” Congo Research Group, `http://congoresearchgroup.org/congolese-election-leaks/`, January 16, 2019.

Table 1: Election Forensics Tollkit: DRC 2019, CENI leaked data

Entity	2BL	LastC	P05s	C05s	DipT	Obs
Turnout	4.216	4.498	0.219	0.197	0.748	15687
	(4.173, 4.261)	(4.456, 4.543)	(0.213, 0.226)	(0.191, 0.203)	–	
X1001_44_10	4.031	3.41	0.393	0.273	0	15687
	(3.962, 4.097)	(3.367, 3.457)	(0.385, 0.401)	(0.267, 0.28)	–	
X1001_48_14	4.11	4.176	0.282	0.234	0	15687
	(4.06, 4.155)	(4.129, 4.224)	(0.275, 0.29)	(0.227, 0.241)	–	
X1001_84_158	4.118	4.276	0.226	0.206	0.039	15687
	(4.07, 4.167)	(4.232, 4.319)	(0.22, 0.233)	(0.199, 0.212)	–	

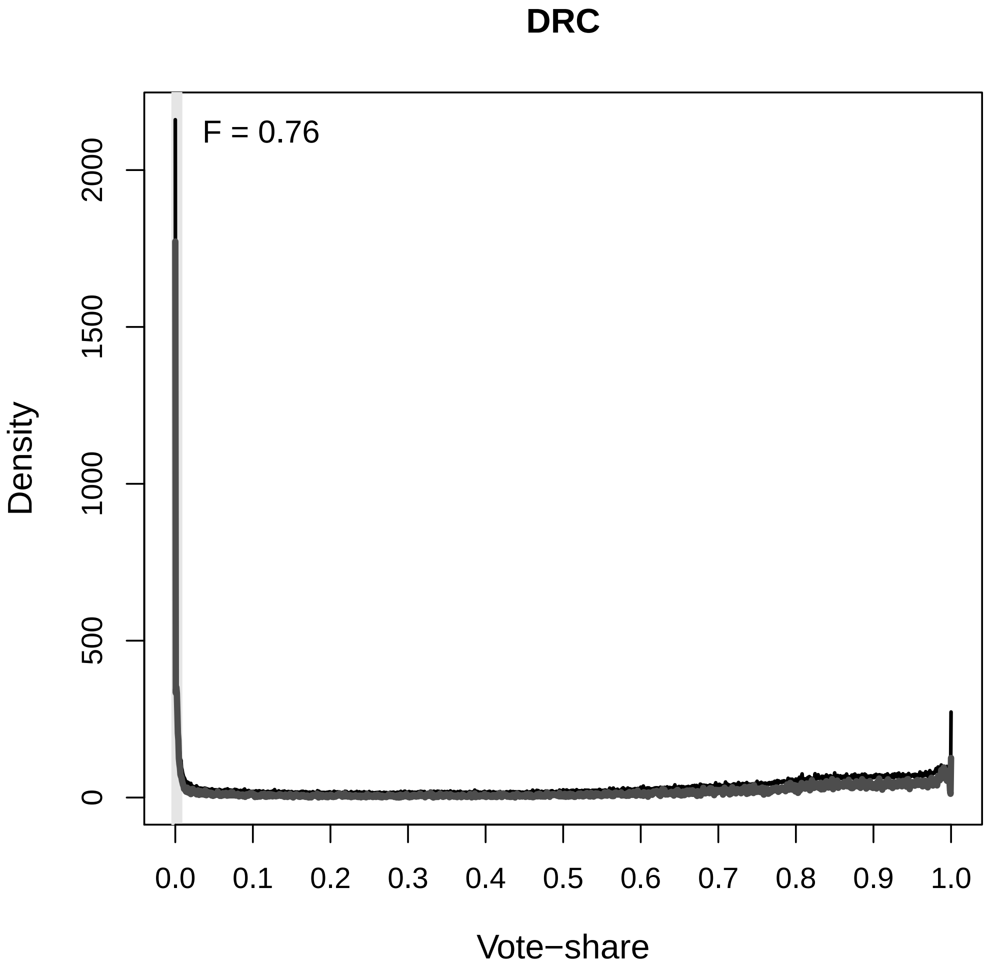
Note: "2BL," second-digit mean; "LastC," last-digit mean; "C05s," mean of variable indicating whether the last digit of the vote count is zero or five; "P05s," mean of variable indicating whether the last digit of the rounded percentage of votes for the referent party or candidate is zero or five; "DipT," p-value from test of unimodality; "Obs," number of polling station observations. Values in parentheses are nonparametric bootstrap confidence intervals.

Table 2: Finite Mixture Model Parameter Estimates

Election	$\hat{f}_{\mathrm{i}}$	$\hat{f}_{\mathrm{e}}$	$\hat{\alpha}$	$\hat{\theta}$	$\hat{\tau}$	$\hat{\nu}$	$\hat{\sigma}_{\tau}$	$\hat{\sigma}_{\nu}$	LR	n
DRC 2019	.120	.000131	1.40	.661	.415	.761	.0582	.141	14855.2	15,317

Note: f_{i} is the probability of incremental fraud, f_{e} is the probability of extreme fraud, α is the stealing/manipulation parameter,θ is the standard deviation of the fraudulent vote proportion, τ is true mean turnout, ν is mean leader's true vote proportion, σ_{τ} is turnout proportion standard deviation, σ_{ν} is leader's true vote proportion standard deviation, LR is the likelihood ratio test statistic for the hypothesis that there are no frauds (i.e., that $f_{\mathrm{i}} = f_{\mathrm{e}} = 0$). n is the number of polling site observations.

Figure 1: Spikes Test, DRC 2019 CENI Leaked Data



Note: $n = 15317$ polling site observations analyzed using the method of Rozenas (2017).