Towards Detecting and Measuring Ballot Stuffing*

Dmitriy Vorobyev†

Abstract

This paper proposes a method for detecting electoral fraud in the form of ballot stuffing. As ballot stuffing increases both turnout and the incumbent's vote share in precincts where it occurs, precincts with low reported turnout are more likely to be clean. Information on clean precincts is used to simulate counterfactual data for "infected" precincts, which are then compared to the observed data. The method is applied to the 2006 Finnish presidential elections. The test fails to reject the hypothesis of no ballot stuffing for the original data, but detects artificially imputed 2% fraud. The same test implies that in the 2004 presidential elections in Russia at least 4.7% of the votes were stuffed in favor of the incumbent.

Keywords: Elections, Fraud Detection

* This Research was supported by a grant RRC X-96 from CERGE-EI Foundation under a program of Global Development Network. All opinions expressed are those of the author and have not been endorsed by CERGE-EI or the GDN.

† Center for Economic Research and Graduate Education - Economic Institute, a joint workplace of Charles University in Prague and the Academy of Science of the Czech Republic.

Address: CERGE-EI, P.O. Box 882, Politickych veznu 7, Prague 1, 111 21, Czech Republic.

1. Introduction

Despite its importance, electoral fraud suffers from a relative lack of attention in the academic literature. Probably the main reason for this is the absence of a reliable measure of fraud. Indeed, not only measuring but even detecting fraud is problematic. The existing methods of fraud detection are more qualitative than quantitative, often based on the subjective assessment of electoral transparency and fairness by observers or other participants of the electoral process, and the results they produce may not always be treated as fully reliable. The few attempts to rigorously analyze electoral data for the presence of fraud have usually required a large amount of data, which handicaps efforts to measure fraud, proxy it, or even detect it with some confidence. It further precludes implementing reliable empirical research, which in turn discourages efforts towards a theoretical study of the nature and consequences of electoral fraud.

This paper proposes a statistical mechanism for testing the fairness of elections when the available data are limited. The methodology enables elections to be tested for the presence of electoral fraud in the form of ballot stuffing using official detailed electoral data. The mechanism is applied to test the fairness of the 2004 Russian presidential elections, whose transparency and integrity are often put in doubt, and obtain an estimate of the magnitude of ballot stuffing in Russia.

The paper is organized as follows. The next section discusses existing approaches of detecting fraud. Section 3 presents a methodology that enables testing the fairness of elections based on official electoral results. In section 4 the methodology is applied to several datasets. First, I create artificial electoral data, show that the test fails to reject the null hypothesis of no ballot stuffing, then impute fraud of about 2% and test the data once again, resulting in a strong rejection of the null hypothesis. Second, I perform the same exercise for data on the 2006 presidential elections in Finland. The test cannot reject the hypothesis of fair elections for the original data but rejects the hypothesis, once 3% fraud is imputed. Third, I apply the test to data on the 2004 Russian presidential elections. The hypothesis of no ballot stuffing is strongly rejected in this case, and the test implies that at least 4.5% of the votes had been stolen in favor of the incumbent.

2. Literature Review

The issue of detecting fraud in data is not new and has been a focus of much research for a long time. The general idea behind the majority of fraud detecting statistical techniques is tracing unusual patterns in the observed data that could occur due to fraud. Such techniques have been successfully used for uncovering fraud in a variety of domains ranging from sports betting (e.g., Wolfers, 2006)) and

education (e.g., Jacob and Levitt, 2003) to online auctions (e.g., Pandit et al., 2007) and banking (e.g., Quah and Sriganesh, 2008). Comprehensive reviews of statistical fraud detecting methods and their applications are presented by Bolton and Hand (2002) as well as Phua et al. (2005). A number of recent papers review fraud detecting techniques for specific fields such as telecommunications (Becker et al., 2010), health care (Li et al., 2008) and finance (Sudjianto et al., 2010). From these papers one can see that the design of each method highly depends on the nature of the data and type of expected fraud, generally, all fraud detecting methods may be divided into two main groups: supervised and unsupervised. Methods of the supervised type assume that there are two data samples available for the analysis: the one which is affected by fraud, and the one which is not. In this case, labeling a new data set as clean or fraudulent is essentially a comparison with benchmark samples. When such samples are not available, the unsupervised methods are applied. They do not use benchmark samples and instead look for outliers in an observable sample. Due to the nature of data on elections and frequently changing electoral environments, electoral fraud detecting methods have to be of an unsupervised type.

Attempts to detect fraud in the electoral process used to be rare and not systematic in the academic literature for a long time. Lehoucq (2003) in his comprehensive review of studies on electoral fraud mentions a number of papers that look for traces of electoral fraud in elections in Argentina, Peru, Colombia, England, Ireland, Germany, Spain, Mexico and some Asian countries.[1] However, the majority of this research detects fraud using descriptive evidence such as surveys, interviews and documents; none uses statistical methods. Even though such approaches can provide insight into the presence of electoral fraud in a given election, they require tremendous effort to collect relevant data and at the same time still produce only qualitative and not always objective conclusions.

Recently researchers have started to pay attention to the statistical analysis of electoral data with the aim of detecting electoral fraud. The largest and most rapidly growing stream of academic literature on electoral fraud detection is digit analysis. Specifically, such studies analyze digit patterns in electoral data, trying to identify anomalies that can appear due to fraud.

Beber and Scacco (2008) suggest a methodology based on the idea that people are bad random number generators: if elections are fair, the distribution of insignificant digits (e.g., digits at the third decimal place and furthers) in electoral outcomes (i.e. data on turnouts and vote shares) must be close to uniform, but if there are manual changes in outcomes there must be biases in generating digits. The idea is supported by a statistical comparison of outcomes from Swedish and Nigerian elections. However, such method can only detect manipulations with electoral returns but cannot produce any result if electoral outcomes are shaped in a more sophisticated way than just manually changing digits in election protocols.

[1] For the references see Lehoucq (2003)

In contrast, a number of recent papers analyze the first significant digits in official electoral data and try to find deviations from Benford's Law (Benford, 1938). The law states that the first digits in real-life data are distributed in a special non-uniform way. Deviations from the law are found by Roukema (2009) in data on the last Iranian presidential elections, Cantu and Saiegh (2010) in Argentinean elections, Pericchi and Torres (2004) in Venezuela, Mebane (2006, 2010) in the US and Mexico and by Breunig and Goerres (2011) in Bundestag elections in Germany. Despite their merits, the methods of digital analysis are subject to criticism which puts in doubt their relevance for detecting fraud in electoral data (e.g., Brady, 2005). Recently Deckert et al. (2011) have shown that deviation from Benford's Law can arise in electoral data regardless of whether the elections are rigged or fair, and moreover the the methods of such type in terms of efficiency in marking elections as clean or fraudulent essentially do not differ from a random draw. Thus, deviations from Benford's Law do not inevitably signal about presence of fraud, and there is still a need in fraud detecting methods.

A number of authors suggest alternative methods of discovering fraud in official electoral data. Myagkov and Ordeshook (2008) examine the fairness of Russian federal elections between 1993—2007, comprehensively studying a variety of patterns in Russian electoral data such as turnout distributions across regions and precincts and vote flows between different elections. They conclude that ballot stuffing as well as other fraud techniques of the 1990s were frequently used in a few Russian ethnic republics but then spread to other regions of the country. They apply similar techniques to study several elections in Russia and Ukraine for the presence of fraud (see Myagkov et al, 2009, for detailed discussion).

Mebane and Kalinin (2009) also study data on Russian federal elections in 2003—2008 using several methods. In addition to the digital tests for deviations from Benford's Law, they analyze turnout distribution and correlation between turnout and the candidates' vote shares to find anomalies that are likely to arise due to fraud.

Electoral fraud in Russia is also discussed by Treisman (2009) who comprehensively reviews the trends in voting in Russia since 1991. In a chapter devoted to electoral manipulations and fraud, by studying a variety of Russian electoral statistics in a way similar to Myagkov and Ordeshook (2008) and Mebane and Kalinin (2009), the author finds that in the early 1990s, the elections in Russia were almost clean, while since 2000 electoral irregularities have become an integral part of electoral competition. Although their approaches are reasonable and capable of producing reliable conclusions, they are mainly based on a visual analysis and comparison of electoral data.

More recently Levin et al (2009) elaborates on the approaches by Myagkov and Ordeshook (2008) and explores electoral data in Venezuela for the presence of fraud. They analyze the data on two consecutive state-level referenda in 2007 and 2009 and, assuming time-constant voters' preferences,

discover unusual patterns in the voting behavior of selected regions that mainly benefit the incumbent. Specifically, the main finding of the paper is that most of the new votes in favor of Chavez in the 2009 referendum came from the regions with large abstention in 2007. To get the result Levin et al. explore three types of indicators. First, they perform digit analysis of the electoral outcomes. Second, they study the flow of votes between the two elections by estimating the proportion of vote share in the first elections that 'flows' to each alternative in the second elections in order to see whether there is a noticeable increase in support of one of the alternatives in regions with substantial increase in turnout. Finally, they closely look at the relationship between the time changes in turnout and share of votes cast for the incumbent. As in Myagkov and Ordeshook (2008), Levin's analysis is primarily based on a comprehensive investigation and description of data patterns rather than statistical testing.

Mainstream election fraud detecting techniques are also presented and discussed in Alvarez (2008, chapter 3). In short, even though electoral fraud appears to be very widespread, existing means of detecting fraud are primarily descriptive and qualitative and highly dependent on the nature of available data. Though statistical studies of electoral data with focus on fraud do exist, they are mainly focused on exploring unusual patterns in data and require a tremendous amount of information. Thus, there is still a need for a rigorous method to detect fraud and measure its extent, especially for cases when available data are restricted. This paper attempts to make a progress towards designing such a method.

3. Fraud Detecting Methodology

The suggested approach is based on the observation that ballot stuffing increases both turnout and votes cast for the corrupt candidate (hereafter I assume that fraud is implemented in favor of the incumbent). If elections are subject to ballot stuffing, which takes place not in all but in a selected number of precincts, this observation immediately implies that a precinct with lower reported turnout is more likely to be clean. The general idea behind the methodology described below is to use the information from such low turnout precincts to simulate counterfactual data for high turnout and likely fraudulent precincts, and test for systematic difference between counterfactual and observed data.

The suggested procedure allows for testing for the presence of ballot stuffing even when very limited data are available. Suppose that the only data available for the analysis are precinct level turnouts and candidates' vote shares. If elections are not fraudulent, and there is a certain degree of homogeneity between electoral precincts, the distribution of turnout across precincts should be close to bell-shaped. Clearly, if electoral districts are similar in terms of characteristics that might determine turnout, or alternatively if turnout is weakly affected by characteristics in which the districts differ, then distribution

of turnout should be approximately normal (see, for example, Myagkov and Ordeshook (2008) or Levin et al. (2009) for more detailed discussion).

In turn, ballot stuffing, when takes place in a given precinct, results in an increase both in reported turnout and in the incumbent's vote share. Consequently, such fraudulent precinct moves in turnout distribution towards its right tail. As a result distribution of turnout in fraudulent elections will be skewed to the left and have thicker right tail. Furthermore, in precinct in the right distribution tail, i.e. those with high reported turnout, incumbent will have an advantage in comparison to the other precincts. The idea behind the suggested methodology is to check whether incumbent has such an advantage and whether it could be considered as natural.

Suppose the following statistic is computed from the available data:

$$s = \frac{V_I/V_R|_{t \geq t^*}}{V_I/V_R|_{t < t^*}}. \quad (1)$$

The nominator of the statistic is the ratio of the incumbent's and the runner-up's shares in the right tail of the turnout distribution (i.e., in those precincts where the turnout is above some threshold t^*). The denominator is the same ratio, but computed over the left distribution tail. I use this statistic to test the null hypothesis that elections are fair. Under the null hypothesis the statistic should be close to one if there is no objective systematic relationship between turnout and voting in favor of one or another candidate, meaning that if elections are fair, the ratio of the incumbent's and the runner-up's shares should not systematically differ in the precincts with high and low turnout. But if there is ballot stuffing in favor of the incumbent, the incumbent's share in the precincts with high turnout will be relatively higher, meaning that the statistic will be greater than one (less than one, if ballot stuffing is in favor of the challenger).

Indeed, there could be an objective correlation between the turnout and the vote shares for the candidates if the supporters of one candidate are more politically active than the supporters of the others. In this case this correlation will be present over the whole dataset, including the left tail. In other words, if such natural relationship between turnout and vote shares exists, it can be estimated using left tail clean data only, and fraud, if it exists, will make this correlation higher in the right tail. The procedure described below is designed to test not for the presence of correlation between turnout and incumbent's vote share in the right tail data, but rather for the presence of extra correlation in comparison to the left tail.

Using the statistic, I test the null hypothesis that there is no ballot stuffing. To conduct the test, one needs to know only a distribution of the statistic under H0. To obtain such distribution the following procedure is proposed.

First, recall that ballot box stuffing increases both turnout and the incumbent's share of votes. This means that a fraudulent precinct "moves" to the right tail of the turnout distribution. This in turn means that in the case of rigged elections, the left tail of the observed turnout distribution contains a fewer number of fraudulent precincts than the right one. Moreover, the higher the scale of fraud, the further to the right a fraudulent precinct moves, implying that a larger tail of the distribution stays clean. Second, recall that if there is no ballot stuffing and precincts are in some sense homogenous, turnout distribution across precincts should be approximately normal. Assuming some particular shape of the true turnout distribution (for example, normal), I choose its parameters such that the distribution fits the left-tail data, i.e. those with turnout below some threshold value t^* (I discuss the choice of the threshold below). Next, I estimate the relationships between turnout and vote shares in the clean left tail by simply regressing vote shares on turnout:

$$V_{Ii} = \alpha + \beta t_i + \varepsilon_i. \qquad (2)$$

Note that the purpose of these regressions is not to establish a causal effect of turnout on vote shares, but rather to find a correlation and then extrapolate it on the simulated right tail.

Then I repeat the following simulation multiple times. At each simulation step I generate a new turnout distribution across precincts $\overline{t_i}$ as a random draw from the fitted normal distribution and "predict" vote shares. To make "predicted" vote shares consistent with clean left tail data I, first, maintain the same relationships between the vote shares and turnout as in the observed clean left tail, and, second, introduce additional uncertainty into predicted vote shares such that their variances evaluated over the left tail are the same as the variances of observed left tail vote shares. Specifically, vote shares for the incumbent and challenger (runner-up) are predicted as

$$\overline{V_{Ii}} = \tilde{\alpha} + \tilde{\beta}\overline{t_i} + u_i, \qquad (3)$$

$$\overline{V_{Ri}} = 1 - \overline{V_{Ii}},$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ are random draws from normal distributions with means α and β, and standard deviations equal to corresponding standard errors from regression (2). The latter means that when predicting the vote share, I do not just use coefficients obtained from regression (2), but allow them to vary across simulations according to the precision of the estimation. If a predicted vote share exceeds 1, it is equalized to 1. Errors u_i are drawn from a zero mean normal distribution. Variance of the distribution is chosen such that

$$Var(\overline{V_I})|_{\bar{t}<t^*} = Var(V_I)|_{t<t^*}.$$

By allowing coefficients to vary and making variance of predicted vote shares to be the same as variance of actual vote shares, I guarantee that the simulated right tail data are consistent with observed relatively clean left tail data.

Once the vote shares are predicted, statistic (1) can be computed. Repeating this simulation multiple times and computing the statistic on each step, one can obtain a distribution of the statistic under the null hypothesis that elections are fair, and tabulate critical values. Computing the statistic for observed data one can now test the hypothesis. Note that if the value of the statistic appears to be above the right tail critical value, it implies that ballot stuffing took place in favor of the incumbent. In contrast, a statistic below the left tail critical value signals ballot stuffing in favor of the challenger. Given critical and actual values of the statistic it is now easy to obtain an estimate of ballot stuffing magnitude by calculating the vote share that incumbent has to obtain in the "right tail" precincts in order to equalize the observed value of the statistic to its critical value of desired confidence level. The difference between actual vote share and counterfactual vote share calculated in this way would give an estimate of ballot stuffing magnitude.

When there are more than two candidates in elections the procedure is slightly different. Because turnout can be related differently to vote shares of each candidate and there are several candidates, challenger's vote share cannot be predicted by simply subtracting incumbent's vote share from one. Instead, his vote share should be obtained in a similar way that incumbent's one (formulas 2 and 3).

The ballot stuffing detection procedure is based on a number of implicit assumptions. First, I assume that ballot stuffing occurs in a "small" number of precincts. Suppose instead that ballot stuffing of relatively the same magnitude would occur in all precincts. This means that turnout and share of votes cast for the incumbent increase in all precincts. Thus, there will be no systematic difference between left and right tail data due to fraud, which is needed for identification. Second, the fraud should be of reasonable magnitude in a sense that it should result in a noticeable increase in turnout to move the precinct to the right tail of the turnout distribution. Together these two assumptions say that fraud, in order to be detected should "move" the precinct where it occurred to the right tail of the distribution.

The methodology described above explicitly distinguishes between the "left tail" and "right" tails of turnout distribution by using a turnout threshold t^*. Ideally, t^* should be chosen such that all precincts with turnout below t^* are clean and the lowest reported turnout among the fraudulent precinct is slightly above t^*. In practice such choice is challenging. More likely, there will still be some fraudulent precincts even in the left turnout distribution tail but less than in the right one, meaning that fraud detection is still possible though the fraud magnitude will be underestimated in this case. On the one hand, the low value of the threshold allows for the capture of small-scale fraud and fraud in low turnout precincts since they are more likely to appear above the threshold. On the other hand, low t^* will not

allow detecting even large-scale fraud if it appears in a very small number of precincts, as the contribution of the fraudulent precincts in the statistic will be relatively small due to a large number of clean precincts. Also, low t^* will result in a small number of data points in the left tail which are used for estimation of the natural relationship between turnout and vote shares which will decrease the power of the test. On the other hand, a high threshold value would make it easier to reject the null hypothesis of no ballot stuffing if there is large-scale fraud, but could fail to reject the hypothesis when fraud is "balanced". Thus, the choice of threshold generally depends on the data as well as some priors about the nature and the extent of fraud.

One way to endogenize the choice of the turnout threshold is to analyze the values of coefficient on turnout from regression (2) for different thresholds. Clearly, going over different threshold value from low to high, at some point the left tail data would start containing fraudulent precincts. As a result coefficient β from regression (2) will start growing if fraud is in favor of incumbent and decreasing if fraud is in favor of challenger. Thus, value of turnout threshold at which coefficient of turnout starts growing (decreasing) would be a natural choice for t^*. The value of t^* specified in this way can itself signal about the nature of fraud. If t^* appears to be high that would mean that in each fraudulent region magnitude of ballot stuffing was huge as substantial share of spoiled precincts ended up in the very right tail of the reported turnout distribution. Alternatively, reasonably low t^* means that fraud in a given precinct was not extreme though total number of fraudulent regions could still be substantial.

Indeed, there could be the cases when sharp changes in the value of β are not observed at all (for instance when elections are clean). In this case the only way to define t^* is to make some reasonable, yet ad-hoc choice, for instance, some number between 0.5 and 0.6.

Noticeable growth or decline of coefficient β starting from some turnout threshold value itself signals about fraud as in the absence of ballot stuffing the coefficient should not sharply change. Thus, testing for the broken trend in β as a function of t^* could be another and probably simpler way to test for the presence of ballot stuffing. Alternatively, because the suggested methodology is based on the observation that turnout in clean elections should follow approximately a normal distribution, one can simply test for the distribution symmetry. However, the suggested methodology has a number of advantages over the alternatives. First, the method would indicate the direction of ballot stuffing if it exists. Depending on whether observed value of the statistic falls to the left or right tail of its distribution under the null hypothesis, one can always say whether ballot stuffing is in favor of the incumbent or challenger. Second and most important, in contrast to the symmetry and broken trend tests, under the suggested method it is possible to obtain an estimate of ballot stuffing magnitude (see section 3.3).

4. Testing Fairness of Elections

In this section I apply the described methodology of detecting ballot stuffing to several distinct datasets. I first generate artificial clean electoral data and then impute fraud into them. I apply the test to original clean data and then to the fraudulent data to show that the test raises a red flag in case of fraudulent data only. Then, I apply the methodology to real data which came from the 2006 Finnish presidential elections. As the integrity of these elections was never subject to debate, I consider them as an example of presumably clean elections and show that the test cannot reject a null hypothesis of no ballot stuffing for this data, but it does reject it once fraud is artificially imputed. Finally, I test for the presence of ballot stuffing the Russian data of the 2004 presidential elections whose integrity is often put in doubt (see section 3.3 for the discussion).

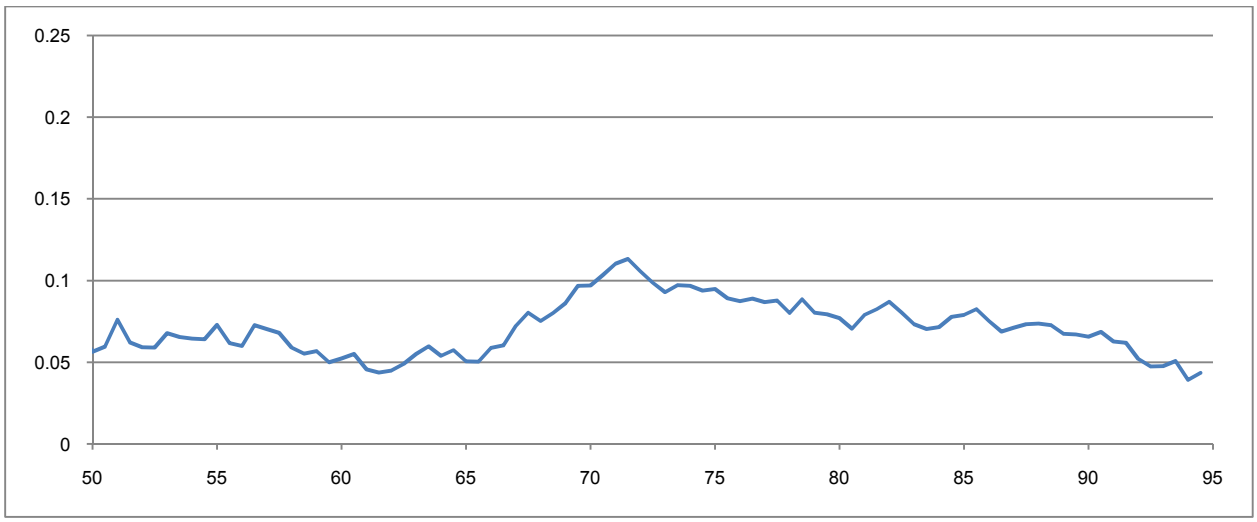
4.1. Artificial Data

First, I show that the method is able to detect electoral fraud of a reasonable magnitude in artificial data. For this purpose I create a dataset that consists of turnout and candidates' vote shares. Specifically, I generate 1000 observations for turnout (t) that follow a normal distribution with 0.5 mean and 0.1 standard deviation. Each observation represents data for a precinct. Then I generate vote shares for the incumbent allowing for a natural correlation between vote shares and turnout as well as noise drawn from normal distribution $N(0, 0.05)$:

$$V_{Ii} = 0.5 + 0.05t_i + \varepsilon_i.$$

I then apply the methodology described above to the simulated data. First, I need to choose a threshold value of turnout to define "left" and "right" tails. Figure 1 shows that there is no clear trend break in coefficient on turnout from regression (2) as a function of turnout percentile where threshold value is evaluated, and the variation in the coefficient is not substantial. I choose threshold value to be at the 61st percentile, which implies that the "left" tail contains precincts with turnout below 0.556.

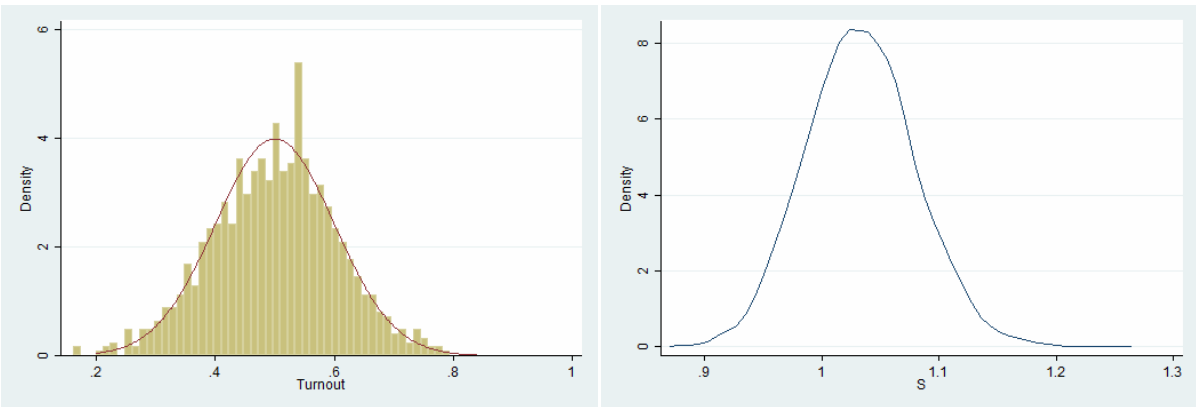
Figure 1. Coefficient on turnout as a function of threshold value for artificial data.



Then I estimate the relationship between the incumbent's vote share in the left tail of the turnout distribution (i.e., in precincts where turnout is less than 0.556) by running a regression of V_I on t. Next, I choose the parameters of the normal distribution such that it fits the left tail of the observed turnout distribution as precisely as possible. Having the regression coefficients, their standard errors and turnout distribution parameters, I predict incumbent vote share in the right tail of the turnout distribution (i.e., in precincts where $t \geq 0.556$) allowing for variation in coefficients (coefficients for prediction are randomly drawn from distributions consistent with the estimated coefficient means and standard errors) and noise. Noise is added in a way that variances of incumbent's predicted and original vote shares are the same for precincts with $t < 0.556$.

Once the right tail data are constructed, the test statistic is calculated. Then I repeat the procedure of the right tail prediction 5000 times, calculate the statistic on each simulation, get the distribution of the statistic under the null hypothesis of no ballot stuffing, tabulate critical values and compare them to the value of the statistic from the observed data. The value of the statistic from the observed data is 1.059, while the 90% critical value is 1.093 and the 10% critical value is 0.973. Thus, the test cannot reject the null hypothesis.

Figure 2. True turnout distribution and distribution of the statistic under the null hypothesis for the original data.



Then, I impute fraud in the data. I randomly choose 150 precincts (15%). In each of them I give the incumbent additional votes: in every spoiled region, the incumbent receives additional number of votes f_i proportional to the size of the district E_i. Then, if we denote $\theta = f_i/E_i$ where E is the size of the district, after fraud turnout $\hat{t}_i$ and incumbent's vote share $\hat{V}_{Ii}$ can be expressed in terms of before fraud turnout t_i and the vote share V_{Ii} as follows:

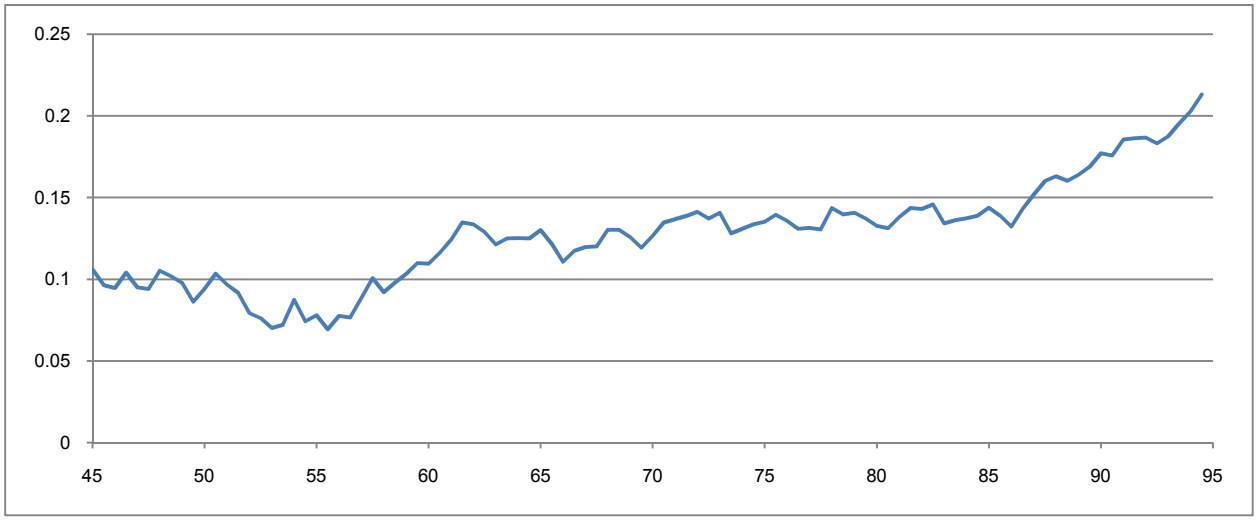
$$\hat{t}_i = t_i + \theta,$$

$$\hat{V}_{Ii} = \frac{V_{Ii}t_i+\theta}{t_i+\theta}.$$

Indeed, the higher θ the easier for the test to reject the hypothesis of no ballot stuffing. Thus, I choose the smallest value of θ such that the null hypothesis is rejected at the 99% confidence level. To guarantee 95% confidence rejection, θ should approximately be 0.18, which gives the incumbent about extra 1.9% of fraudulent votes on aggregate level measured as the difference between his before and after fraud vote shares. Figure 4 shows after fraud distributions of turnout. One can see that fraud results in a thicker right distribution tail.

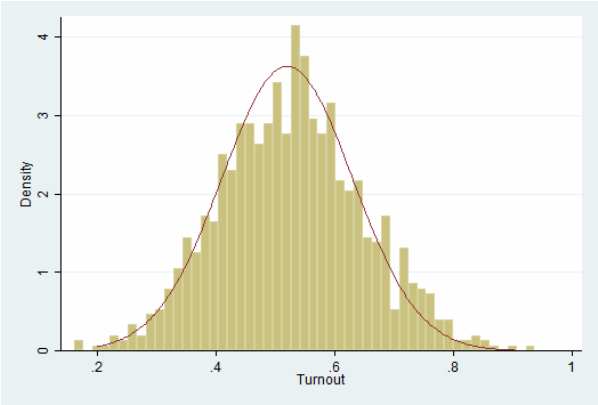
Once fraudulent data are generated I apply the detecting procedure described in section 2. For this test threshold value is chosen to be at 55.5th percentile, where according to Figure 3 coefficient on turnout starts persistently growing.

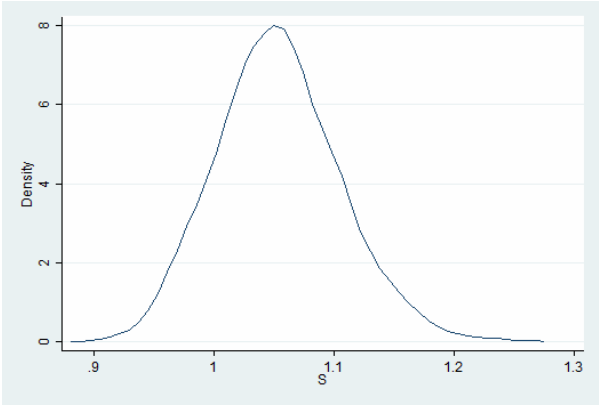
Figure 3. Coefficient on turnout as a function of threshold value for artificial data with imputed fraud.



With $\theta = 0.18$, the value of the statistics is 1.215 and the 99% critical value is 1.185. Figure 4 presents the distribution of the statistic under the null hypothesis of no ballot stuffing. One can see that this distribution is not exactly as the one in Figure 2. This is because the distribution on Figure 5 is obtained using the after fraud data, and, as it was discussed above, the left tail of turnout distribution can still contain fraudulent precincts, which fully account for the observed difference.

Figure 4. Distribution of the statistic under the null hypothesis for fraudulent data.





The exercise shows that the suggested methodology does not reject the hypothesis of fair elections even if there is natural positive correlation between turnout and share of votes for the incumbent but successfully detects with 99% confidence less than 2% fraud.

4.2. Finnish 2006 Presidential Elections

In this section the methodology is applied to real data that came from a presumably clean first round of the 2006 presidential elections in Finland. These elections were chosen as one of the rare examples of direct executive elections whose integrity can be hardly put in doubt. OSCE Office for Democratic Institutions and Human Rights (ODIHR), the largest and probably the most experienced organization that deploys elections observation missions in Europe, was requested to observe the Finnish Parliamentary elections of 2007 held one year after the presidential race. In their report OSCE analysts recommended that "no OSCE/ODIHR election observation or assessment activity shall be undertaken in connection with the 18 March 2007 parliamentary elections. A tradition of democratic elections in Finland is accompanied by a commensurate level of public trust. All interlocutors expressed their overall confidence in the electoral process, and no immediate issues were brought to the attention of the Needs Assessment Mission that would necessitate OSCE/ODIHR involvement." [2] Thus, there is no strong reason to question the integrity of the Finnish presidential elections of 2006.

Another reason why Finnish presidential elections were chosen to test the ballot stuffing detection methodology is that these elections were in some sense close to the Russian presidential elections of 2004, which are analyzed further. Though it is hard to believe that Russian and Finnish elections are truly comparable in any dimension, this is probably the best match one could do: the dates of the elections were not too far apart, the electoral systems in both countries are close, and the importance of elections is in some sense similar: directly elected presidents of Russia and Finland both have an executive power in contrast to the majority of European countries. In fact, only few countries in Europe

[2] Republic of Finland. Parliamentary Elections 18 March 2007, OSCE/ODIHR Needs Assessment Mission Report. Page 4. Available at http://www.osce.org/odihr/elections/finland/24126 (retrieved 18.08.2011).

have a directly elected president as an executive (Armenia, Azerbaijan, Belarus, Bulgaria, Cyprus, Finland, France, Georgia, Lithuania, Poland, Romania, Slovakia, Ukraine) and Finland seems to be the best choice from this sample if one would like to have an example of a country which has as many as possible similarities with Russia in terms of electoral environment and, what is the most important for this paper, the highest confidence in electoral transparency and integrity.

For the analysis I use the data for the first round of the elections. There main reason for such a choice is again an intention to make Finnish elections as comparable as possible to the 2004 Russian elections analyzed further. Because since 1996 the second round in Russian presidential elections has never been held due to the winning of one of the candidates in the first round, only first round Russian data are available for the analysis. Thus, it necessary to analyze first round data also for Finland as the ballot stuffing detection procedure for more than two candidate elections slightly differs from the one applied to the artificial data in the previous section. As it was discussed in section 2, in such case the method requires analysis of the correlation between turnout and vote shares not only for incumbent but also for the challenger.

The dataset consists of 461 municipality-level data observations which came from the Finnish public authority "Statistics Finland".[3] To perform the analysis I first need to choose turnout threshold value. Following the approach suggested in section 2 I draw coefficient on turnout form regression (2) as a function of threshold value. It can be seen from Figure 5 that there is no clear break of the trend which would suggest a choice of the threshold. So I choose two different threshold values at the 56.5th percentile where there is small growth of the coefficient and at the 88th percentile where one can see a small decline in the graph.

In both cases the test does not reject the hypothesis of no ballot stuffing: with turnout threshold value chosen at the 56.5th percentile the statistic is 0.685, while the 90% and the 10% critical values are 0.801 and 0.467 respectively, and for the 88th percentile threshold the value of statistic is 0.601, while the 90% and the 10% critical values are 0.733 and 0.422 respectively. Note that a critical value below 1 suggests that natural correlation between turnout and votes for the incumbent is lower than between turnout and votes for runner-up.

Figure 5. Coefficient on turnout as a function of threshold value for original Finnish data.

[3] www.tilastokeskus.fi

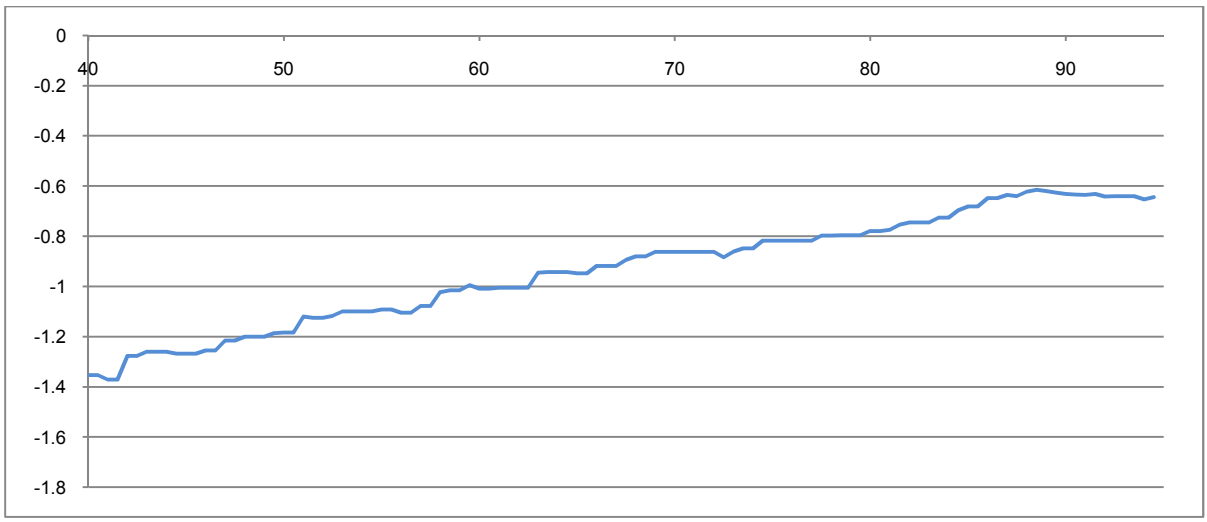
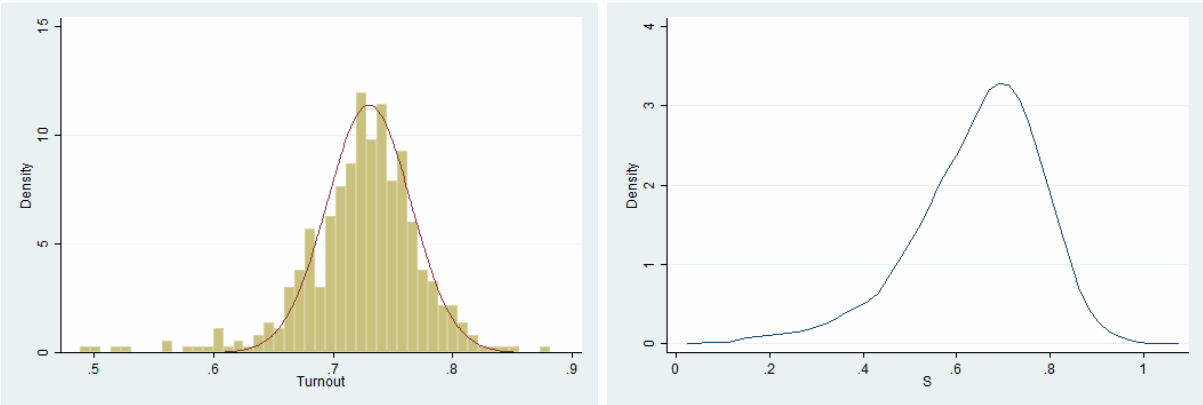
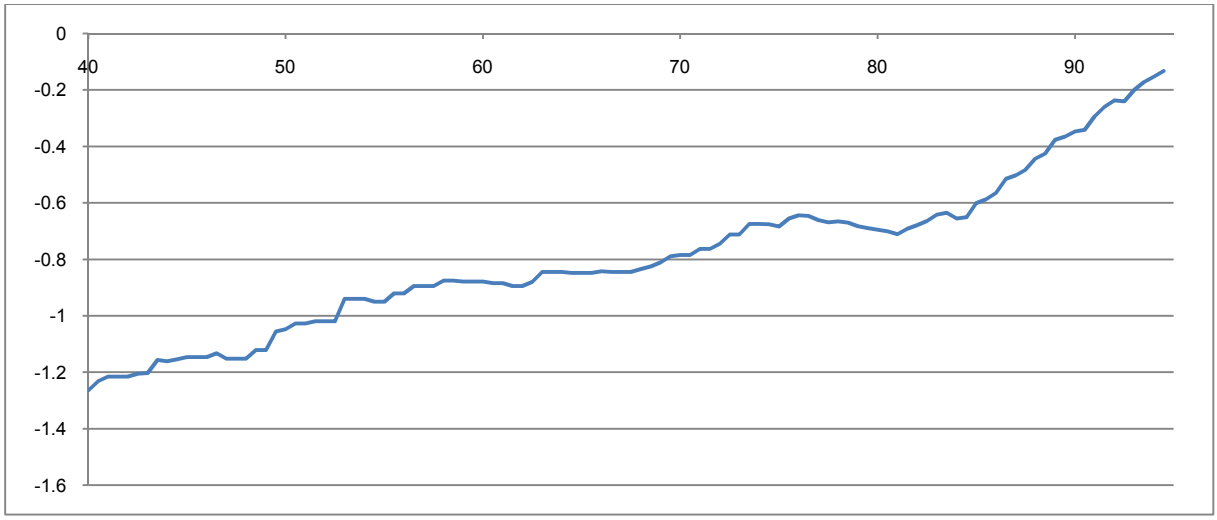


Figure 6. Turnout distribution and distribution of the statistic under H0 for the original data and 56.5 threshold.



Then, as in the previous case, I randomly choose 69 precincts (15%), artificially "spoil" them by adding additional 20% of votes in favor of the incumbent which gives him approximately extra 2.5% of votes, and conduct the test once again. Now the threshold value is chosen at the 81th percentile.

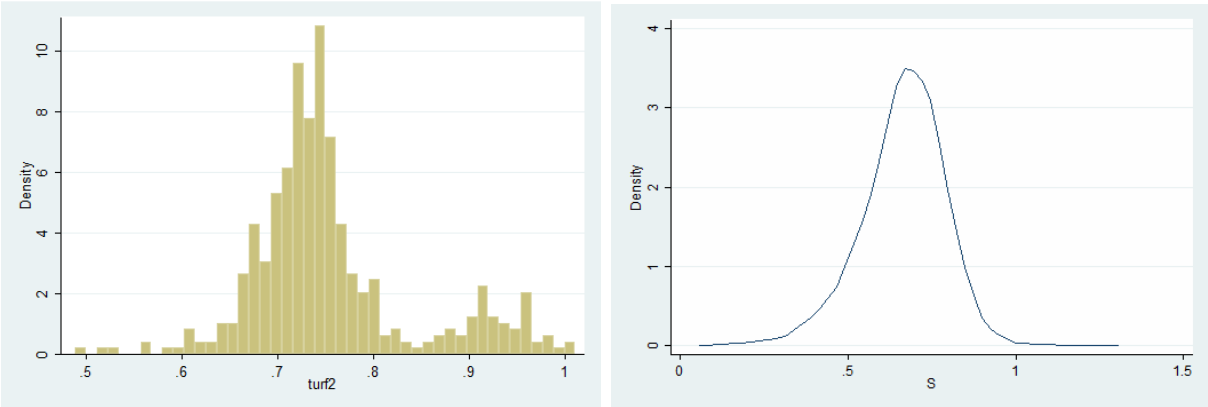
Figure 7. Coefficient on turnout as a function of threshold value for Finnish data with imputed fraud.



With these 2.5% fraudulent data, the test rejects the null hypothesis of fair elections at the 99% confidence level: statistic 1.018, 99% critical value – 0.928. Counterfactual victory margin in this case

with 99% confidence should not exceed 23.86% while observed victory margin is 24.74% and before fraud one (true) is 22.24%. Thus, in this case the method is unable to detect less than 1.6% ballot stuffing. As it was discussed in section 2 underdetection is explained by the fact that left tail of the turnout distribution may still contain a certain amount of the

Figure 8. After fraud turnout distribution and the distribution of the statistic under H0 for fraudulent data.



4.3. Russian 2004 Presidential Elections

Finally, I apply the methodology to the Russian presidential elections of 2004 which are often considered to be not fair (Treisman (2009), Myagkov and Ordeshook (2008), Sakwa (2005)). The data are obtained from the central elections commission of Russia, and contain information about the number of registered voters, turnout and votes cast for each candidate for each precinct (2670 observations) in the first round.

The results of the test support the conventional wisdom regarding the fairness of the elections: with turnout threshold value taken at the 69th percentile (Figure 9) the value of the statistics is 1.763, while the 99% critical value is 0.948. Note that such critical value says that correlation between the turnout and votes for the incumbent must be lower than correlation between turnout and votes for the challenger, but in the reported data it is much higher.

Figure 9. Coefficient on turnout as a function of threshold value for 2004 Russian Presidential Elections.

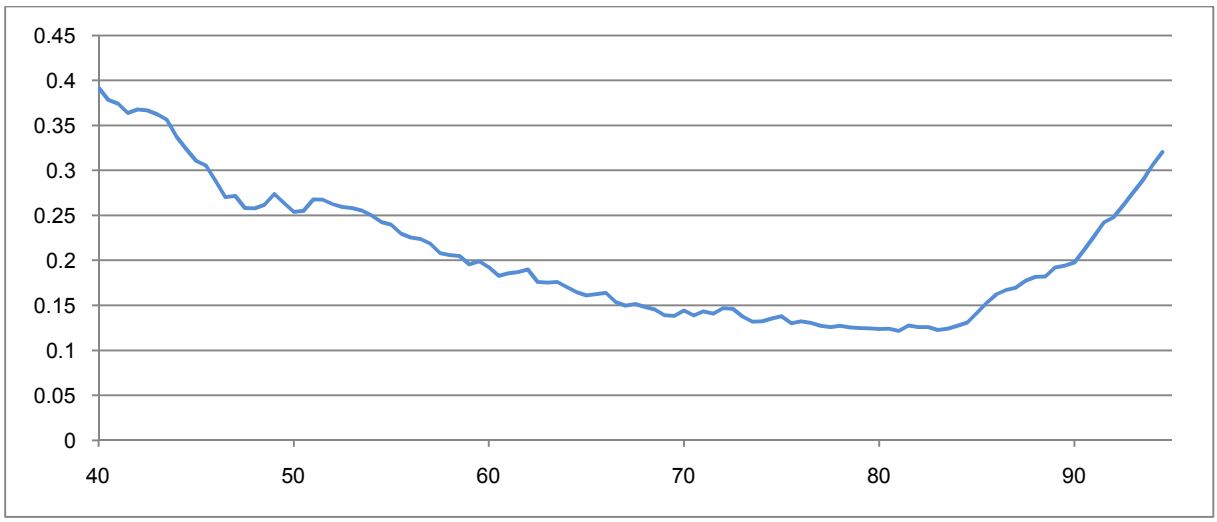
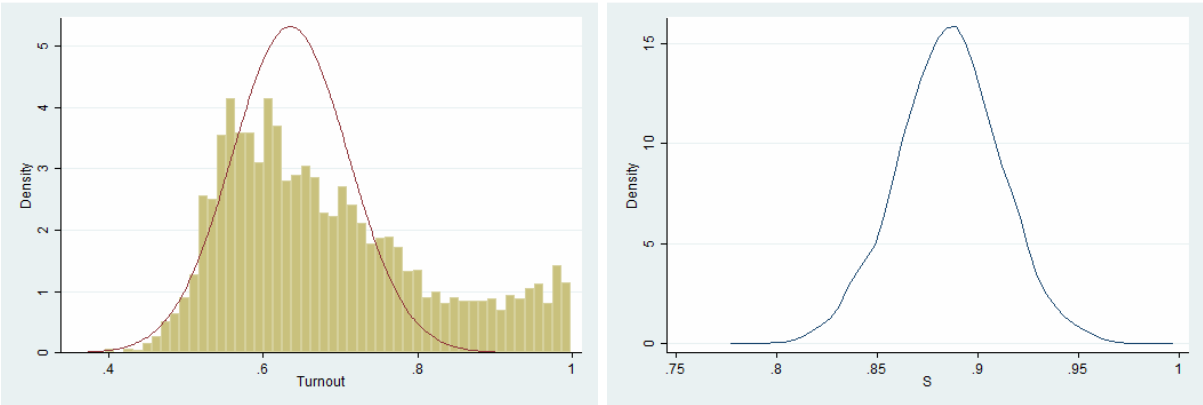


Figure 10. Turnout distribution and the distribution of the statistic under the null hypothesis



Finally, using the difference between critical and actual values of the statistic I can get an estimate of the fraud magnitude. Given that the 99% critical value for the statistic is 0.948 victory margin should not exceed 54% while the observed victory margin is 58.64%. Thus, with 99% confidence in 2004 Russian presidential elections incumbent Vladimir Putin got at least extra 4.64% of votes, which corresponds to more than 7 000 000 stuffed ballots. However, even if corrected for ballot stuffing, the incumbent's vote share would exceed 50% that would still allow him to win in the first round. The estimate could be thought of a low bound of electoral fraud because, first, the methodology itself underestimates ballot stuffing as it was discussed in section 2, and second, it detects ballot stuffing only, while the set of rigging techniques is indeed wider.

5. Conclusion

This paper suggests a simple statistical method for testing elections for the presence of ballot stuffing using official detailed electoral data. The method is based on the observation that ballot stuffing increases both turnout and the incumbent's vote share in precincts where it occurs. Hence, precincts with relatively low reported turnout are more likely to be clean. Using the information on relatively

clean precincts, it is possible to simulate counterfactual data for “infected” precincts and compare them with the observed data.

The method is first piloted on artificial and artificially fraudulent real data, and subsequently applied to test the fairness of the Russian executive elections in 2004, whose transparency and integrity are dubious. Results strongly reject the hypothesis of no ballot stuffing and suggest that at least 7 000 000 ballots in 2004 Russian elections were stuffed in favor of the incumbent, which however did not alter the winner of the elections.

References

Alvarez, M., Hall, T., Hyde, S., 2008. Election Fraud: Detecting and Deterring Electoral Manipulation, *Brookings Institution Press,* Washington, D.C.

Bolton, R., Hand, D., 2002. Statistical Fraud Detection: A Review. *Statistical Science,* Vol. 17, No. 3, 235–255.

Benford, F., 1938. The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, Vol. 78, 4, 551–572.

Berber, B., Scacco, A., 2008. What the Numbers Say: a Digit Based Test for Election Fraud Using New Data from Nigeria. *Presented at the Annual Meeting of the APSA*.

Becker, R., Volinsky, C., Wilks, A., 2010. Fraud Detection in Telecommunications: History and Lessons Learned. *Technometrics*, 52(1), 20–33.

Brady, H., 2005. Comments on Benford s Law and the Venezuelan Election. *Mimeo*.

Breuning, C., Goerres, A., 2011. Searching for Electoral Irregularities in an Established Democracy: Applying Benford's Law Tests to Bundestag Elections in Unified Germany. *Electoral Studies*, forthcoming.

Cantu, F., Saiegh, S., 2010. A Supervised Machine Learning Procedure to Detect Electoral Fraud Using Digital Analysis. *Working Paper*, University of California.

Decker, J., Myagkov, M., Ordeshook, P., 2011. Benford's Law and the Detection of Election Fraud. *Political Analysis,* 19 (3), 245-268.

Jacob, B., Levitt, S., 2003. Rotten Apples: An Investigation of The Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics,* August, 843–877.

Lehoucq, F., 2003. Electoral Fraud: Causes, Types and Consequences. *Annual Review of Political Science*, 6, 233–256.

Li, J., Huang, K., Jin, J., Shi, J., 2008. A Survey on Statistical Methods for Health Care Fraud Detection. *Health Care Management Science*, Vol. 11, 3, 1386–9620.

Lukinova, E., Myagkov, M., Ordeshook, P., 2011. Ukraine 2010: Were Tymoshenko's Cries of Fraud Anything More Than Smoke? *Post-Soviet Affairs*, 27, 1, 37–63.

Mebane W., 2006. Election Forensics: Vote Counts and Benford's Law. *Working Paper,* University of Michigan.

Mebane W., 2010. Election Fraud or Strategic Voting? *Presented at the Annual Meeting of the MPSA.*

Mebane W., Kalinin, K., 2009. Comparative Election Fraud Detection. *Presented at the Annual Meeting of the APSA.*

Myagkov, M., Ordeshook, P., 2008. Russian Elections: An Oxymoron of Democracy. *Working Paper*, California Institute of Technology.

Myagkov, M., Ordeshook, P., Shakin, D., 2009. The Forensics of Election Fraud: Russia and Ukraine. Cambridge University Press.

Pandit, S., Chau, D., Wang, S., Faloutsos, C., 2007. NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks. *Proceedings of the 16th international conference on WWW.*

Phua, C., Lee, V., Smith, K., Gayler, R., 2005. A Comprehensive Survey of Data Mining-Based Fraud Detection Research. *Artificial Intelligence Review*.

Pericchi, L., Torres, D., 2004. La Ley de Newcomb-Benford y sus aplicaciones al Referendum Revocatorio en Venezuela. *Reporte Tecnico No-Definitivo*, 2a.

Quash, J., Sriganesh, M., 2008. Real-Time Credit Card Fraud Detection Using Computational Intelligence. *Expert Systems with Applications*, Vol. 35, 4, 1721–1732.

Roukema, B., 2009. Benford’s Law Anomalies in the 2009 Iranian Presidential Election. *Mimeo*.

Sakwa, R., 2005. The 2003-2004 Russian Elections and Prospects for Democracy. *Europe-Asia Studies*, Vol. 57, No. 2, 369–398.

Sudjianto, A.,Nair, S., Yuan, M., Zhang, A., Kern, D., Cela-Díaz, F., 2010. Statistical Methods for Fighting Financial Crimes. *Technometrics*, 52(1), 5–19.

Treisman, D., 2009. Elections in Russia, 1991—2008. *Working paper*, Moscow State University — Higher School of Economics.

Wolfers, J., 2006. Point Shaving: Corruption in NCAA Basketball. *American Economic Review*, Vol. 96, 2, May, 279–283.