Election Forensics: Outlier and Digit Tests in America and Russia*

Walter R. Mebane, Jr.†

May 1, 2008

Abstract

I illustrate election forensic testing using a combination of robust overdispersed multinomial model estimation and second-digit mean testing, with data from the 2004 U.S. presidential election in Ohio and the 2004 Russian presidential election.

*Prepared for presentation at The American Electoral Process conference, Center for the Study of Democratic Politics, Princeton University, May 1–3, 2008. I thank Kirill Kalinin for helpful advice and assistance.

†Professor, Department of Political Science and Department of Statistics, University of Michigan, Haven Hall, Ann Arbor, MI 48109-1045 (Phone: 734-763-2220; Fax: 734-764-3522; E-mail: wmebane@umich.edu).

Introduction

Two methods that have figured prominently in work on election forensics are robust estimation of count data models with outlier detection (Wand, Shotts, Sekhon, Mebane, Herron, and Brady 2001; Mebane and Sekhon 2004; Mebane and Herron 2005) and tests based on the so-called second-digit Benford's Law (2BL) distribution applied to vote counts (Mebane 2006a,b, 2007b,a, forthcoming). Mebane (2007a) sharpens the question of whether departures from the 2BL distribution are useful for detecting gross departures from an untainted voting process, asking whether the particular pattern of departure is useful for diagnosing precisely why anomalies may have happened. The analysis in Mebane (2007a) ultimately focuses on the conditional means of the second digits in collections of vote counts, measuring how these means differ from the means expected according to the 2BL distribution. The conditioning factors in that analysis, which examined data from the 2006 election in Mexico, were the partisan affiliations of mayors in Mexican municipalities.

The focus of the current paper is to illustrate such conditional digit-mean analysis where the conditioning factors come from robust estimation of the distribution of measures of plausible methods by which an election may have been corrupted. Following a brief simulation exercise intended to suggest intuition, I look at an example from the 2004 U.S. presidential election in Ohio and at an example from the 2004 presidential election in Russia.

2BL Test Statistics

I use two kinds of statistics to implement tests of whether vote counts have the 2BL distribution. One is the Pearson chi-squared statistic used in Mebane (2006a) and subsequent work. To define this statistic, let q_j denote the expected relative frequency, according to Benford's law, with which the second significant digit is j.[1] Let n_j be the number of times the second digit is j among the

[1] The expected frequencies are (rounded) $(q_0, \ldots, q_9) = (.120, .114, .109, .104, .100, .097, .093, .090, .088, .085)$.

vote counts being considered, and define $N = \sum_{j=0}^{9} n_j$. The statistic for a 2BL test is

$$X_{2BL}^2 = \sum_{j=0}^{9} \frac{(n_j - Nq_j)^2}{Nq_j} \,.$$

This statistic may be compared to the χ^2-distribution with 9 degrees of freedom (χ_9^2).

The other kind of test, illustrated in this paper, compares the arithmetic mean of the second digits to the mean value expected if the digits are 2BL-distributed. This test adapts an idea used in Grendar, Judge, and Schechter (2007)'s analysis that focuses on the first significant digit and is intended to identify what they describe as generalized Benford distributions. Grendar et al. suggest that data that do not conform to Benford's law may have first digits that match a member of a specified class of exponential families. In Mebane (2006b) I argue that vote counts in general do not have digits that match Benford's law at all. For instance, the distribution of the first digits of vote counts is undetermined. Mebane (2006b) demonstrates a pair of naturalistic models that produce simulated vote counts with second digits but not first digits that are distributed roughly as specified by Benford's law. Nonetheless we can use the mean of the second digits to test how closely the digits match the 2BL distribution. Given 2BL-distributed counts, the value expected for the second-digit mean is (rounded) $\sum_{j=0}^{9} jq_j = 4.187$.

This second kind of test straightforwardly supports asking whether the distribution of second digits differs from the nominal 2BL distribution in a way that depends on observed conditioning factors. For vote counts y_i observed for precincts or polling stations indexed by i, it's simply a matter of regressing (by ordinary least squares) the second digits on a set of conditioning factors X_i and testing whether the associated coefficients are statistically distinguishable from zero. When the number of counts in the analysis is large, simple normal theory methods are at least a reasonable first-cut method for making inferences.

Simulations

First let me illustrate what happens to the second-digit distributions when votes are artificially manipulated, first in a proportionally uniform manner and then nonuniformly in some simulations.

Start with a situation in which there is an ideal set of numbers that perfectly satisfy the 2BL distribution, and imagine that all the numbers are grossly increased to such an extent that the second digits all change. Table 1 shows what happens in such circumstances. Each row of Table 1 shows an imagined ordering for the second digits d_j, indexed by $j = 0, \ldots, 9$, followed by the implied 2BL mean value $\sum_{j=0}^{9} d_j q_j$. The first row shows the unshifted digits, the second row shows the digits shifted up by one (e.g., 0 becomes 1, 1 becomes 2, and so forth), continuing through the tenth row which shows the digits shifted up by nine. Another perspective on these shifts is to recognize that the digits in the second row occur if a set of 2BL-distributed counts are all increased by ten percent, the digits in the third row occur if the increase is by twenty percent, and so forth. Yet another perspective is that the digit ordering in the second row occurs if all the digits are shifted down by nine; for suitably large counts, this corresponds to decreasing all the original counts by ninety percent. In every case in these nine rows of the table, such uniform shifts in the digits produces increases the 2BL mean. Perhaps counterintuitively, then, the mean of the second digits in a set of counts should be greater than the 2BL expectation of 4.187 if what was originally a set of 2BL-distributed values has been either uniformly increased or decreased by a sufficiently large proportion.

*** Table 1 about here ***

The last three rows of Table 1 illustrate that 2BL means less than the 2BL expectation can occur if changes away from the original 2BL distribution occur in a nonuniform way.

Next consider a simulation in which votes are manipulated nonuniformly. To simulate vote counts for a pair of candidates, I apply a mechanism used in Mebane (2007c). For a set of N precincts indexed by i, compute a set of uncorrelated bivariate normal pseudorandom numbers (denoted (x_i, y_i)) and a set of numbers uniformly distributed on the interval from zero to one (denoted r_i):

$$
\begin{aligned}
(x_i, y_i) &\sim N(\mu_x, \mu_y; \sigma_x, \sigma_y) \\
r_i &\sim U(0,1)\,.
\end{aligned}
$$

The parameters μ_x and μ_y denote the means of x and y, and σ_x and σ_y denote their variances. The (x_i, y_i) values are used to generate proportions of support for each candidate in precinct i:

$$
\begin{aligned}
p_{xi} &= \frac{\exp(x_i)}{\exp(x_i) + \exp(y_i) + 1} \\
p_{yi} &= \frac{\exp(y_i)}{\exp(x_i) + \exp(y_i) + 1}\,.
\end{aligned}
$$

The p_{xi} and p_{yi} values are the proportions of voters in precinct i who vote, respectively, for each candidate. Notice that $p_{xi} + p_{yi} < 1$: the mechanism accommodates ballots that lack a vote for either candidate. To get the number of votes for each candidate we fix a maximum number of potential votes in each precinct, denoted M, so that $\lfloor M r_i \rfloor$ corresponds to the number of ballots cast in precinct i. The simulated counts of votes for the candidates are

$$
\begin{aligned}
z_{xi} &= \lfloor M r_i p_{xi} \rfloor \\
z_{yi} &= \lfloor M r_i p_{yi} \rfloor\,.
\end{aligned}
$$

I simulate counts for pairs of candidates with $M = 2500$ and, successively, $N \in \{500, 1000, 2000\}$.[2] The manipulations consist of a fixed number fM of votes being added to one candidate's total in each precinct and subtracted from the precinct totals of the other candidate; if the number subtracted from the "donor" candidate exceeds the count originally simulated for that candidate, then zero is assigned for that candidate's vote in that instance. I simulate switching for proportions $f \in \{0, .05, .1, .15, .2, .25, .3, .35, .4, .45, .5\}$.

Table 2 shows that the second-digit means for these simulated votes sometimes exceed but

[2]The other simulation parameters are set at values $\mu_x = 1.25$, $\mu_y = 1.0$; $\sigma_x = 1.25$, $\sigma_y = 1.5$.

mostly are smaller than the 2BL expected value. The top part of the table shows the simple difference between the second-digit means and the 2BL expected value 4.187, averaged over the replications. The bottom part of the table shows t-statistics produced by dividing these differences by the standard deviation of the means across replications. For the "receiver" candidate the second-digit means are significantly less than 4.187 for switching proportions $f \in \{.2, .25, .3, .35, .4, .45\}$. Interestingly, for $f = .5$ the "receiver" candidate's second-digit means are significantly greater than 4.187. Differences for the "donor" candidate are not significant in the set of simulation conditions reported here.

*** Table 2 about here ***

Votes and Turnout in Ohio 2004

Several well-known manipulations attempted in Ohio during the 2004 presidential election involved voter turnout. First, Ohio Secretary of State Ken Blackwell attempted to suppress turnout by imposing unreasonable and illegal requirements on the kind of paper to be used for voter registration forms (House Judiciary Committee Democratic Staff 2005). Second, inadequate provision of voting machines produced great delays in voting at many polling places that deterred many people from voting (e.g. Mebane 2005). Whether the latter problem reflected intentional efforts to reduce turnout is not clear, but the fact that turnout was reduced is not in dispute. Table 3 reproduces a display from Mebane and Herron (2005) that shows turnout to be strongly affected by the number of voting machines per registered voter in each precinct, particularly in counties that used direct record electronic (DRE) or punchcard machines.

*** Table 3 about here ***

The table reports coefficients from estimating a set of robust overdispersed multinomial regression models. To assess whether precincts that had unusual levels of turnout also had unusual vote distributions, I use a function of the weights produced as part of the robust estimation algorithm as a regressor in a linear regression model for the second digits of the vote counts recorded for John Kerry and for George W. Bush. These weights vary from zero to one, and they are smaller when the turnout in a precinct is highly disrepant from the value expected given the regressors and coefficients shown in Table 3. As the footnote in the table indicates, overall there are 61 precincts that are outliers, which means the weights for those precincts are zero. Other precincts have weights that are greater than zero but still less than one. Denoting the weights by w_i, the regressor I use in the model for the second digits of the vote counts is $X_i = 1 - w_i$. If the coefficient for this variable is significantly different from zero, the suggestion is that unusual turnout is associated with unusual vote totals for at least one of the candidates.

Table 4, which reports the regression results, shows that over all the precincts in Ohio, the second digit mean when turnout is not unusual relative to the models of Table 3 (i.e., where $w_i = 1$) differs slightly but significantly from the 2BL expected value. The intercepts in both the model for Bush and the model for Kerry are significantly greater than 4.187. In the model for Bush but not in the model for Kerry, the coefficient for the $1 - w_i$ variable is significantly negative. The 95% confidence interval for the mean of the second digit in Bush vote counts when $w_i = 0$ is $(3.45, 3.96)$, which is strictly less than the 2BL expected value.

*** Table 4 about here ***

This significant result for Bush lines up with the result for the "Receiver" candidate in the simulation exercise reported in Table 2. The suggestion is that the substantial turnout manipulations in Ohio in 2004 favored Bush. The fact that the second-digit means for both Bush and Kerry exceed the 2BL expected value in precincts that did not have unusual turnout relative to the Table 3 models also has an important implication. The results of Table 1 suggest that such an elevation in the second-digit means is a symptom of a relatively uniform shift in all the vote counts. Because voting machine shortages and other adminstrative failures in Ohio in 2004 reduced turnout relatively equally among both Democrats and Republicans (Mebane and Herron 2005; Voting Rights Institute 2005, Section III), such a relatively uniform shift is more likely a prevalent reduction in the number of votes for each candidate than a general pattern of increases. So we might conclude that while overall votes were lost in a similar, relatively uniform manner for both candidates, Bush gained votes in those precincts where turnout was unusual relative to expectations based on the provision of voting machines and proportion of the precinct's population that was African-American.

Votes and Manipulations in Russia 2004

Elections in Russia have long been suspect. Based on analysis focused on aberrant patterns in voter turnout, Myagkov and Ordeshook (2008) argue that during the past 15 years "falsifications in the form of stuffed ballot boxes and artificially augmented election counts" have become prevalent throughout the country (see also Myagkov, Ordeshook, and Shaikin forthcoming). The OSCE identified serious problems in the 2004 election (OSCE Office for Democratic Institutions and Human Rights 2004), and by 2008 problems had become so severe that international observer groups declined to observe the election (OSCE Office for Democratic Institutions and Human Rights 2008). Kalinin (2008) demonstrates several significant distortions in votes reported for the 2004 and 2008 presidential elections.

Here I reexamine a few of the hypotheses Kalinin suggests about fraud in Russian elections, using data from the 2004 presidential election. The 2004 election ultimately included six candidates, nominally representing six political parties or coalitions: Vladimir Vladimirovich Putin (United Russia); Nikolaj Mihajlovich Haritonov (Communist, Agrarian); Irina Mucuovna Hakamada (Union of Right); Oleg Aleksandrovich Malyshkin (Liberal Democratic); Sergej Mihajlovich Mironov (Life); and Sergej YUr'evich Glaz'ev (Rodina). In 2004 ballots also offered the option of voting to reject all the candidates, with a choice labeled *Protiv vseh* (Against all). My analysis is based on vote counts reported at the level of polling stations or UIKs (UIK is *uchastkovaya izbiratelnaya komissiya*).[3] The data include counts from 95,426 UIKs, covering 2,755 territories and 89 regions.

Vladimir Putin was universally recognized to have an extremely strong position in the election, both because of his high popularity and because he controlled the apparatus

[3] Data were downloaded on February 16, 2008, from the website of the Central Election Commission of the Russian Federation, `http://www.vybory.izbirkom.ru/region/izbirkom`.

administering the election. Of the hypotheses that Kalinin (2008) proposes regarding the means by which frauds may have been perpetrated in the election, I consider three: H1, artificially elevated voter turnout (the phony votes supposedly going to Putin); H4, excess numbers of invalid or lost votes (votes for opponents to Putin supposedly being spoiled); H5, excess numbers of absentee certificates (certificates supposedly being forged and also used to force voters to cast coerced votes for Putin). The data I use to measure the key variables the hypotheses involve vary slightly from the implementations used by Kalinin (2008). Total turnout I measure using the number of nonabsentee valid ballots at each UIK. The invalid/lost ballot count is the number of invalid or lost ballots. The residual category—nonvoted ballots—is the number of eligible voters at each UIK minus the total of nonabsentee valid ballots plus invalid or lost ballots plus absentee certificates.

There is a question whether electoral manipulations were organized at the level of regional or territorial election commissions. To investigate this, my analytical strategy is to start by estimating four-category robust overdispersed multinomial models separately for the set of regional totals and the set of territorial totals. The four categories are nonabsentee valid ballots, invalid or lost ballots, absentee certificates and nonvoted ballots. In these models the only regressors are the intercepts for each category, so outliers, if any, will be regions or territories that have proportions in a category that are far greater or far smaller than the generally prevailing rate. Table 5 reports the estimates for these intercept parameters along with the number of zero weights estimating the model produces. Note that with four categories of counts, a single observation may include up to three zero weights. Table 6 shows the breakdown by category not only of the numbers of zero weights but also the numbers of weights less than one. Measured in regions or territories, the greatest number both of outliers and of downweighted counts occur for the H5 (absentee certificates) category. The H1 (nonabsentee valid ballots) category has the next highest number, and the H4 (invalid or lost votes) category has the fewest. In terms of the number of UIKs in the unusual regions or territories, however, the H1 category has by far the largest number of unusual observations (e.g., more than 20,000 UIKs in outlier regions or outlier territories), with the H5 category second largest. With one exception—one of the territory outliers for the nonabsentee valid ballots counts—all of the outliers correspond to counts greater than expected according to the generally prevailing rates.

*** Tables 5 and Table 6 about here ***

The next step is to regress the second digits of the UIK-level vote counts for each candidate (and Against All) on region and territory regressors defined as $X_{hi} = 1 - w_{hi}$ where w_{hi} is the weight for category $h \in \{\text{H1}, \text{H4}, \text{H5}\}$ for the region or territory that include the UIK indexed by i. Before considering these regression results, it is worthwhile to take a quick look at the raw relationship between the UIK-level turnout, invalid and absenteee proportions and both the raw vote counts and the vote proportions for each candidate. Figures 1 and 2 respectively show these relationships for candidates Putin and Haritonov. The scatterplot in each subfigure displays a one percent random sample of the full set of UIKs, while the line drawn in each figure represents the linear least squares regression line estimated using all the UIKs. In Figure 1 the vote proportions for Putin show a strong pattern of increase as the turnout, invalid and absenteee proportions increase, but no such pattern is apparent among Putin's raw vote counts. The picture for Haritonov, in Figure 2, also shows increases in the vote proportions, although the relationships are less steep than the ones observed for Putin. The raw vote counts for Haritonov are if anything more negatively associated than Putin's are with the turnout, invalid and absenteee proportions. Such plots suggest Putin's support tends to increase more than other candidates' support with increases in the turnout, invalid and absenteee proportions, although of course such relationships of themselves are in no way diagnostic of anomalies or frauds. The relationships with the raw vote counts are not readily interpretable, but they are worth keeping in mind given our focus on regressions in which the dependent variable is the second significant digit in each of these counts.

*** Figure 1 and Figure 2 about here ***

Table 7 shows an example of the second-digit regression results. The dependent variable in this example is the second digit of the UIK vote counts for Putin. The second digit mean when turnout is not unusual is slightly but significantly less than the 2BL expected value. In light of the simulation results reported in Table 2, this suggests there are nonuniform but pervasive distortions in the vote counts for Putin. All three of the coefficients for the region-level X_{hi} variables are significant, while only one of the coefficients for a territory-level X_{hi} variable is (marginally) significant. This suggests that regions that have outlying levels of turnout exhibit especially distorted Putin vote totals but outlying territories do not. Here, however, it is important to notice that the regional X_{H1i} coefficient is positive while the regional X_{H4i} and X_{H5i} coefficients are negative. Plus the X_{H1i} coefficient is an order of magnitude smaller than the other two. Hence if we considers the mean second digit in UIKs that have $w_{hi} = 0$, the mean when $w_{H1i} = 0$ is closer to the 2BL expected value (although still significantly smaller), while the mean when $w_{H4i} = 0$ or $w_{H5i} = 0$ is even smaller than the overall mean is.

*** Table 7 about here ***

Figures 3 through 9 present the second-digit regression results for all of the candidates in graphical form. Figure 3 plots the point estimates for the intercept parameters for each candidate, with lines to show the 95% confidence interval for each estimate. All but the estimate for the Against All (None in the figure) category are significantly less than the 2BL expected value of 4.187 (indicated with a vertical line in the figure).

*** Figures 3 about here ***

Figure 4 shows the coefficient point estimates and 95% confidence intervals for the X_{H1i} (turnout) regressors. The regional coefficients for Putin and for Malyshkin are significantly positive, and the region coefficient for Mironov and both coefficients for Against All (None) are significantly negative. It is perhaps intuitive that in regions and territories with outlying turnout levels the counts for the Against All alternative are distorted. Figure 5 shows the point estimates and 95% confidence intervals for the mean of the second digit in each candidate's vote counts when $w_{H1i} = 0$; these intervals are all strictly less than the 2BL expected value.

*** Figure 4 and Figure 5 about here ***

Figure 6 shows that the coefficient estimates for the regional X_{H4i} (invalid/lost) regressors are significantly positive for Putin and Hakamada and significantly positive for Mironov and Against

All (None). None of the territorial coefficients are significant. Figure 7 shows that the mean of the candidate vote count second digits when the regional $w_{H4i} = 0$ is significantly less than the 2BL expected value for the first four candidates and significantly greater than the expected value for Mironov and Against All (None). The mean is significantly less than the 2BL expected value when the territorial $w_{H4i} = 0$ only for Haritonov and Mironov.

*** Figure 6 and Figure 7 about here ***

Figure 8 shows that the coefficient estimates for the regional X_{H5i} (absentee) regressors are significantly negative for all the candidates except Malyshkin, while none of the territorial X_{H5i} coefficients are significant. Figure 9 shows that the mean of the vote count second digits when the regional $w_{H4i} = 0$ is significantly less than the 2BL expected value is significantly less than the expected value for all the candidates except Malyshkin.

*** Figure 8 and Figure 9 about here ***

For two of the three classes of possible manipulations the hypotheses focus on, the distribution of votes (not second digits) in categories defined by the outlier weights w_{hi} supports a conclusion that the manipulations did help Putin and harm the other candidates. Table 8 reports the proportion of votes for each candidate among sets of UIKs defined by either $w_{hi} = 1$ or $w_{hi} = 0$. For H1 and H4, Putin's vote proportions are much larger when $w_{hi} = 0$ than when $w_{hi} = 1$, while for most of the other candidates—and in particular for the Communist party candidate Haritonov—the vote proportions are smaller when $w_{hi} = 0$. This pattern does not hold for H5.

*** Table 8 about here ***

Conclusion

Combining robust count model estimation with second-digit testing in relation to the 2BL distribution produces suggestive results worthy of further investigation. For Ohio 2004, an examination of aberrations related to voter turnout suggests that against a background of pervasive distortion in the vote counts for both Bush and Kerry, extremes in unusual turnout were associated with advantages for Bush. For Russia 2004, the analysis suggests advantages for Putin related to both unusual turnout and excessive numbers of invalid and lost ballots. Findings relating to possible manipulation via absentee certificates are more complicated. For both elections, the kind of election forensic testing illustrated here match allegations supported by both contemporaneous observers and other forms of analysis.

A principal direction for further development of the methods themselves is to investigate further the relationship between second-digit means and patterns of shifts in votes. Whether the second-digit mean is greater or smaller than the 2BL expected value does not so far appear to be strongly connected to whether votes have been artificially increased or decreased, but this is a matter for further research.

References

Grendar, Marian, George Judge, and Laura Schechter. 2007. "An Empirical Non-Parametric Likelihood Family of Data-Based Benford-Like Distributions." *Physica A* forthcoming.

House Judiciary Committee Democratic Staff. 2005. "Preserving Democracy: What Went Wrong in Ohio." URL http://www.house.gov/judiciary_democrats/ohiostatusrept1505.pdf.

Kalinin, Kirill. 2008. "Electoral Frauds in Russia: 2004 and 2008 Presidential Elections Compared." Paper presented at the 2008 Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 3.

Mebane, Walter R., Jr. 2005. "Voting Machine Allocation in Franklin County, Ohio, 2004: Response to U.S. Department of Justice Letter of June 29, 2005." Unpublished MS.

Mebane, Walter R., Jr. 2006a. "Detecting Attempted Election Theft: Vote Counts, Voting Machines and Benford's Law." Paper prepared for the 2006 Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 20–23.

Mebane, Walter R., Jr. 2006b. "Election Forensics: Vote Counts and Benford's Law." Paper prepared for the 2006 Summer Meeting of the Political Methodology Society, UC-Davis, July 20–22.

Mebane, Walter R., Jr. 2007a. "Election Forensics: Statistics, Recounts and Fraud." Paper presented at the 2007 Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 12–16.

Mebane, Walter R., Jr. 2007b. "Evaluating Voting Systems To Improve and Verify Accuracy." Paper presented at the 2007 Annual Meeting of the American Association for the Advancement of Science, San Francisco, CA, February 16, 2007, and at the Bay Area Methods Meeting, Berkeley, March 2, 2007.

Mebane, Walter R., Jr. 2007c. "Statistics for Digits." Paper presented at the 2007 Summer Meeting of the Political Methodology Society, Pennsylvania State University, July 18–21.

Mebane, Walter R., Jr. forthcoming. "Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections." In R. Michael Alvarez, Thad E. Hall, and Susan D. Hyde, editors, *The Art and Science of Studying Election Fraud: Detection, Prevention, and Consequences*, .Washington, DC: Brookings Institution.

Mebane, Walter R., Jr. and Michael C. Herron. 2005. "Ohio 2004 Election: Turnout, Residual Votes and Votes in Precincts and Wards." In Democratic National Committee Voting Rights Institute, editor, *Democracy at Risk: The 2004 Election in Ohio*, .Washington, D.C.: Democratic National Committee.

Mebane, Walter R., Jr. and Jasjeet S. Sekhon. 2004. "Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data." *American Journal of Political Science* 48 (Apr.): 392–411.

Myagkov, Misha and Peter C. Ordeshook. 2008. "Russian Elections: An Oxymoron of Democracy." Caltech/MIT Voting Technology Project, VTP Working Paper #63, Mar 2008.

Myagkov, Misha, Peter C. Ordeshook, and Dimitry Shaikin. forthcoming. "Estimating the Trail of Votes in Russia's Elections and the Likelihood of Fraud." In R. Michael Alvarez, Thad E. Hall, and Susan D. Hyde, editors, *The Art and Science of Studying Election Fraud: Detection, Prevention, and Consequences*, .Washington, DC: Brookings Institution.

OSCE Office for Democratic Institutions and Human Rights. 2004. "Russian Federation Presidential Election, 14 March 2004: OSCE/ODIHR Election Observation Mission Report." Warsaw, June 2, 2004.

OSCE Office for Democratic Institutions and Human Rights. 2008. "OSCE/ODIHR Regrets that Restrictions Force Cancellation of Election Observation Mission to Russian Federation." Press release, Warsaw, February 7, 2008.

Voting Rights Institute, Democratic National Committee. 2005. *Democracy at Risk: The 2004 Election in Ohio*. Washington, D.C.: Democratic National Committee.

Wand, Jonathan, Kenneth Shotts, Jasjeet S. Sekhon, Walter R. Mebane, Jr., Michael Herron, and Henry E. Brady. 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* 95 (Dec.): 793–810.

Table 1: Effect of Shifting Second Digits on 2BL Digit Means

digit order	2BL mean
0 1 2 3 4 5 6 7 8 9	4.187390
1 2 3 4 5 6 7 8 9 0	4.337416
2 3 4 5 6 7 8 9 0 1	4.461716
3 4 5 6 7 8 9 0 1 2	4.558196
4 5 6 7 8 9 0 1 2 3	4.624448
5 6 7 8 9 0 1 2 3 4	4.657676
6 7 8 9 0 1 2 3 4 5	4.654594
7 8 9 0 1 2 3 4 5 6	4.611298
8 9 0 1 2 3 4 5 6 7	4.523083
9 0 1 2 3 4 5 6 7 8	4.384182
0 1 2 3 4 5 6 7 8 8	4.102392
0 1 2 3 4 5 6 7 7 8	4.014822
0 1 2 3 4 5 6 7 7 7	3.929825

Notes: For digits d_j in the displayed ordering $j = 0, \ldots, 9$, and 2BL probabilities q_j, the displayed 2BL mean is $\sum_{j=0}^{9} d_j q_j$.

Table 2: Simulated Vote Switching: Observed Mean vs. 2BL Mean

Second-digit Mean Minus 4.187

fraction	Receiver (cand. 1) 500	Receiver (cand. 1) 1000	Receiver (cand. 1) 2000	Donor (cand. 2) 500	Donor (cand. 2) 1000	Donor (cand. 2) 2000
0	−.06	−.07	−.07	−.06	−.07	−.06
0.05	.01	.01	.01	−.07	−.07	−.06
0.1	.05	.04	.04	−.07	−.06	−.06
0.15	−.06	−.07	−.07	−.07	−.07	−.08
0.2	−.33	−.32	−.32	−.14	−.12	−.10
0.25	−.20	−.19	−.20	−.12	−.14	−.13
0.3	−.22	−.22	−.23	−.14	−.13	−.13
0.35	−.43	−.44	−.43	−.12	−.13	−.12
0.4	−.93	−.94	−.94	−.03	−.09	−.05
0.45	−.35	−.34	−.33	.00	.02	.01
0.5	.17	.17	.18	.09	.09	.10

t-statistic

fraction	Receiver (cand. 1) 500	Receiver (cand. 1) 1000	Receiver (cand. 1) 2000	Donor (cand. 2) 500	Donor (cand. 2) 1000	Donor (cand. 2) 2000
0	−0.5	−0.8	−1.1	−0.5	−0.7	−1.0
0.05	0.1	0.1	0.2	−0.5	−0.6	−0.8
0.1	0.4	0.4	0.6	−0.4	−0.6	−0.7
0.15	−0.4	−0.7	−1.2	−0.4	−0.6	−0.9
0.2	−2.5	−3.5	−4.8	−0.7	−0.8	−1.0
0.25	−1.5	−2.3	−3.3	−0.6	−0.9	−1.3
0.3	−1.7	−2.6	−3.6	−0.6	−0.8	−1.0
0.35	−3.2	−4.8	−6.8	−0.4	−0.7	−0.9
0.4	−7.0	−11.3	−14.4	−0.1	−0.4	−0.3
0.45	−3.2	−4.3	−5.6	0.0	0.1	0.1
0.5	1.5	2.3	3.4	0.2	0.3	0.5

Notes: Simulated second-digit mean minus the 2BL expected value 4.187, averaged over 250 Monte Carlo replications. Fraction denotes the proportion of votes shifted from the donor candidate to the receiver candidate. The top number in each column shows the number of precincts used in the referent simulation.

Table 3: Voter Turnout: Machine Technology, Machines per Voter and Precinct Racial Composition Regressors

Variable	DRE			Punchcard		
	Coef.	SE	t-ratio	Coef.	SE	t-ratio
(Intercept)	0.26	0.0318	8.17	0.754	0.0221	34.1
Machines per Registered Voter	74.60	7.0100	10.60	38.600	2.6900	14.3
Proportion African American	−0.98	0.0438	−22.40	−0.851	0.0380	−22.4
	Optical Central			**Cuyahoga**		
Variable	Coef.	SE	t-ratio	Coef.	SE	t-ratio
(Intercept)	0.976	0.0432	22.60	0.630	0.0289	21.80
Machines per Registered Voter	23.500	6.3300	3.71	−10.100	3.2100	−3.13
Proportion African American	−0.689	0.0545	−12.70	−0.371	0.0201	−18.50
	Optical Precinct			**Hamilton**		
Variable	Coef.	SE	t-ratio	Coef.	SE	t-ratio
(Intercept)	0.783	0.0976	8.020	0.212	0.167	1.27
Machines per Registered Voter	−5.770	14.3000	−0.402	117.000	17.500	6.67
Proportion African American	−2.360	0.2630	−8.940	−0.610	0.044	−13.90

Notes: Robust ($\tanh$) overdispersed binomial regression estimates. For each precinct or ward, the dependent variable counts the number of registered voters voting versus the number of registered voters not voting. DRE precincts: LQD $\sigma = 4.82$; $\tanh$ $\sigma = 4.66$; $n = 1,535$; 7 outliers. Optical Central precincts: LQD $\sigma = 3.91$; $\tanh$ $\sigma = 3.92$; $n = 807$; 6 outliers. Optical Precinct precincts: LQD $\sigma = 3.08$; $\tanh$ $\sigma = 3.11$; $n = 139$; 1 outlier. Punchcard precincts: LQD $\sigma = 4.51$; $\tanh$ $\sigma = 4.26$; $n = 5,478$; 28 outliers. Cuyahoga precincts: LQD $\sigma = 3.67$; $\tanh$ $\sigma = 3.53$; $n = 1,411$; 15 outliers. Hamilton precincts: LQD $\sigma = 4.14$; $\tanh$ $\sigma = 4.10$; $n = 979$; 4 outliers. Punchcard precincts exclude Cuyahoga and Hamilton precincts.

Table 4: Candidate Vote Second Digits Regressed on Voter Turnout Weights, Ohio 2004

Variable	Bush			Kerry		
	Coef.	SE	t-ratio	Coef.	SE	t-ratio
(Intercept)	4.360	0.0291	149.6	4.410	0.0287	153.6
One Minus Turnout Weight	−0.649	0.2260	−2.9	0.134	0.2220	0.6

Notes: Ordinary least squares regression estimates. For each precinct or ward, the dependent variable is the second significant digit of the respective candidate's vote count. Bush: RMSE = 2.87; $n = 10,240$. Kerry: RMSE = 2.84; $n = 10,336$.

Table 5: Ballots in the 2004 Russian Presidential Election

	Regions			Territories		
Category	Coef.	SE	t-ratio	Coef.	SE	t-ratio
Nonabsentee Valid Ballots (H1)	-4.04	0.057	-70.9	-4.12	0.0147	-281
Invalid or Lost Ballots (H4)	0.373	0.0178	21.0	0.397	0.00479	82.9
Absentee Certificates (H5)	-4.21	0.0475	-88.6	-4.23	0.0137	-308
Nonvoted Ballots	—	—	—	—	—	—

Notes: Robust ($\tanh$) overdispersed binomial regression estimates. The dependent variable is the count for each category in each UIK. The only regressor in the model is the intercept for each category, estimates for which appear in the table, with Nonvoted Ballots being the reference category. Regions: LQD $\sigma = 53.8$; $\tanh$ $\sigma = 46.5$; $n = 89$; 20 zero weights. Territories: LQD $\sigma = 13.6$; $\tanh$ $\sigma = 11.7$; $n = 2,755$; 648 zero weights.

Table 6: Russia 2004 Outlier Weight Statistics

Regions

	Regions			Included UIKs		
weights	H1	H4	H5	H1	H4	H5
$w < 1$	2	1	42	28,547	2,921	2,559
$w = 0$	1	0	19	23,892	0	3,214

Territories

	Territories			Included UIKs		
weights	H1	H4	H5	H1	H4	H5
$w < 1$	157	69	1,425	29,777	2,670	3,837
$w = 0$	42	4	602	20,264	146	1,358

Notes: w denotes $\tanh$ weights from the four-category robust overdispersed multinomial model for constant proportions (intercept only). $w < 1$ if the absolute studentized residual $|r| \geq 1.88$, $w = 0$ if $|r| > 4$.

Table 7: Russia 2004, Putin Vote Counts' Second-digit Regression Estimates

Category	Coef.	SE	t-ratio
(Intercept)	4.0816	$.015$	279.3
X_{H1} Region	0.0666	$.023$	2.9
X_{H1} Territory	0.0454	$.023$	1.9
X_{H4} Region	-0.6166	$.205$	-3.0
X_{H4} Territory	0.0568	$.137$	0.4
X_{H5} Region	-0.5260	$.054$	-9.8
X_{H5} Territory	0.0374	$.065$	0.6

Notes: Ordinary least squares regression estimates. The dependent variable is the second significant digit in the vote count for Putin in each UIK. RMSE $= 2.92$; $n = 95,129$.

Table 8: Russia 2004, Vote Proportions by Outlier Weight Categories

	Region Weight Categories					
	H1		H4		H5	
Candidate	$w=1$	$w=0$	$w=1$	$w=0$	$w=1$	$w=0$
Putin	$.411$	$.612$	$.452$	—	$.459$	$.410$
Haritonov	$.083$	$.082$	$.089$	—	$.091$	$.044$
Hakamada	$.027$	$.020$	$.025$	—	$.023$	$.049$
Malyshkin	$.012$	$.013$	$.013$	—	$.014$	$.007$
Mironov	$.005$	$.005$	$.005$	—	$.005$	$.004$
Glaz'ev	$.027$	$.027$	$.027$	—	$.026$	$.038$
Against All	$.024$	$.020$	$.023$	—	$.021$	$.039$
	Territory Weight Categories					
	H1		H4		H5	
Candidate	$w=1$	$w=0$	$w=1$	$w=0$	$w=1$	$w=0$
Putin	$.441$	$.532$	$.457$	$.497$	$.458$	$.440$
Haritonov	$.089$	$.088$	$.089$	$.037$	$.089$	$.075$
Hakamada	$.026$	$.022$	$.025$	$.036$	$.025$	$.029$
Malyshkin	$.013$	$.013$	$.013$	$.007$	$.013$	$.011$
Mironov	$.005$	$.005$	$.005$	$.003$	$.005$	$.004$
Glaz'ev	$.028$	$.022$	$.027$	$.028$	$.026$	$.031$
Against All	$.023$	$.020$	$.022$	$.025$	$.022$	$.025$

Notes: Number of votes for each candidate in each category divided by the number of eligible voters.

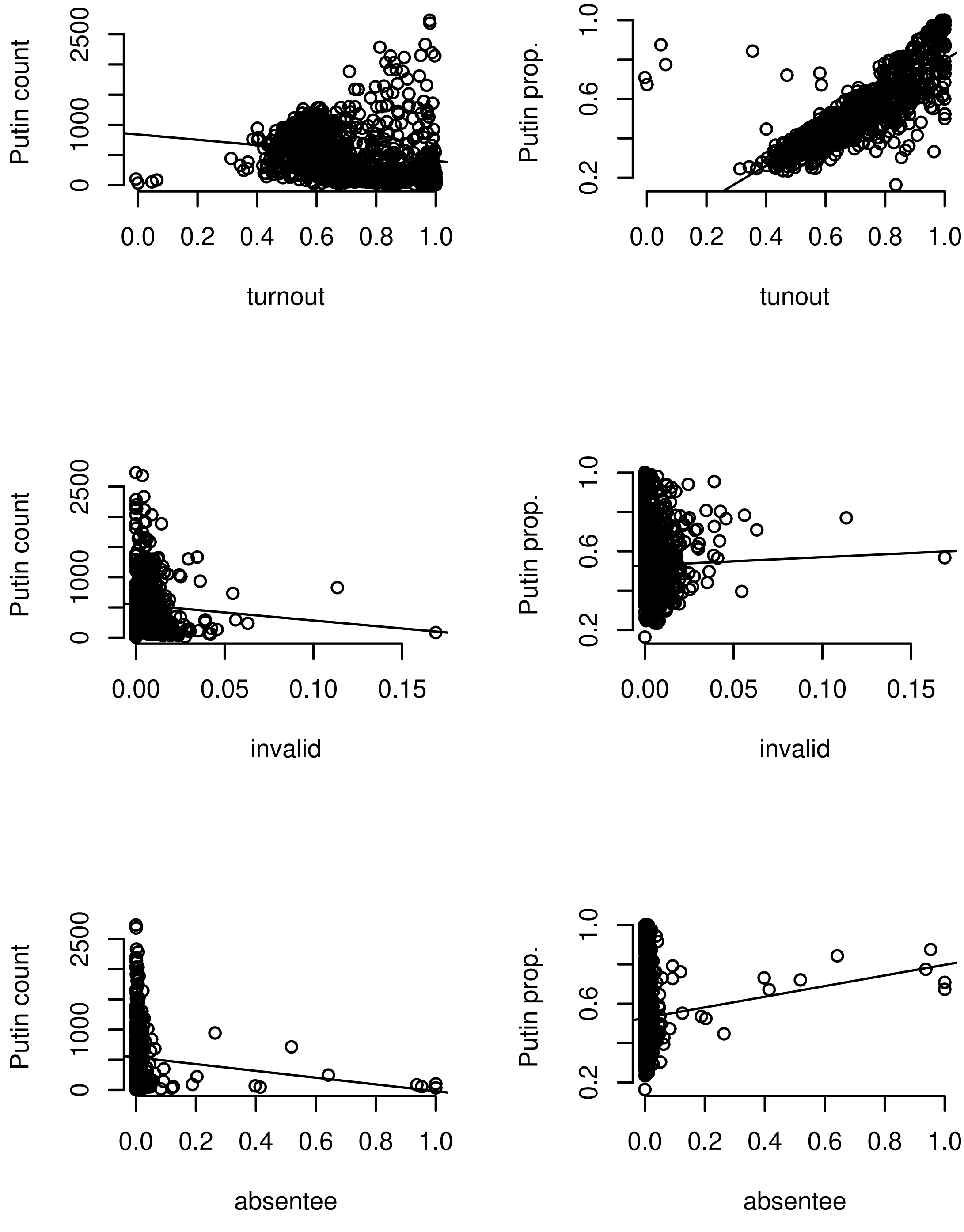

Figure 1: Vote counts and proportions (one percent sample), 2004: Putin

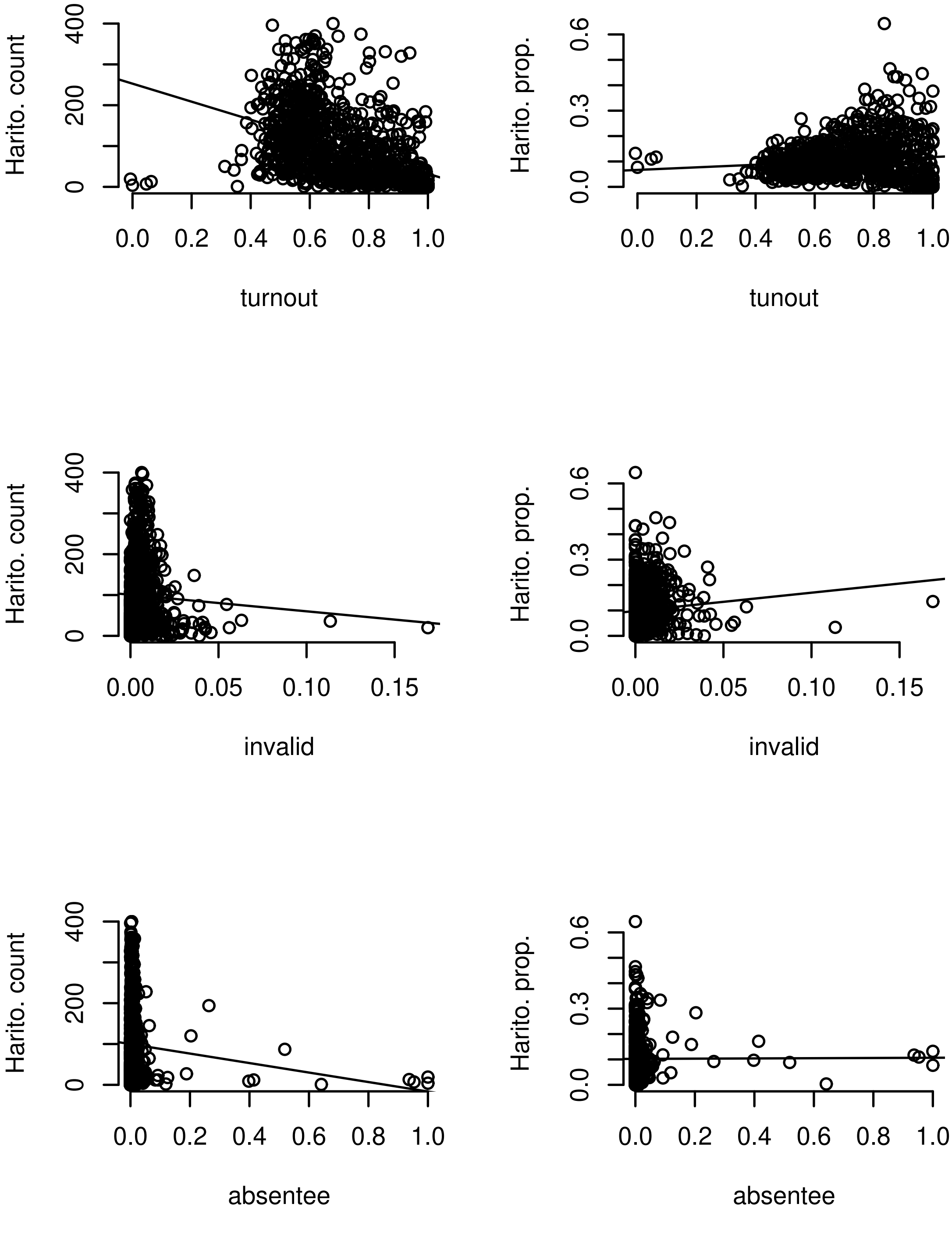

Figure 2: Vote counts and proportions (one percent sample), 2004: Haritonov

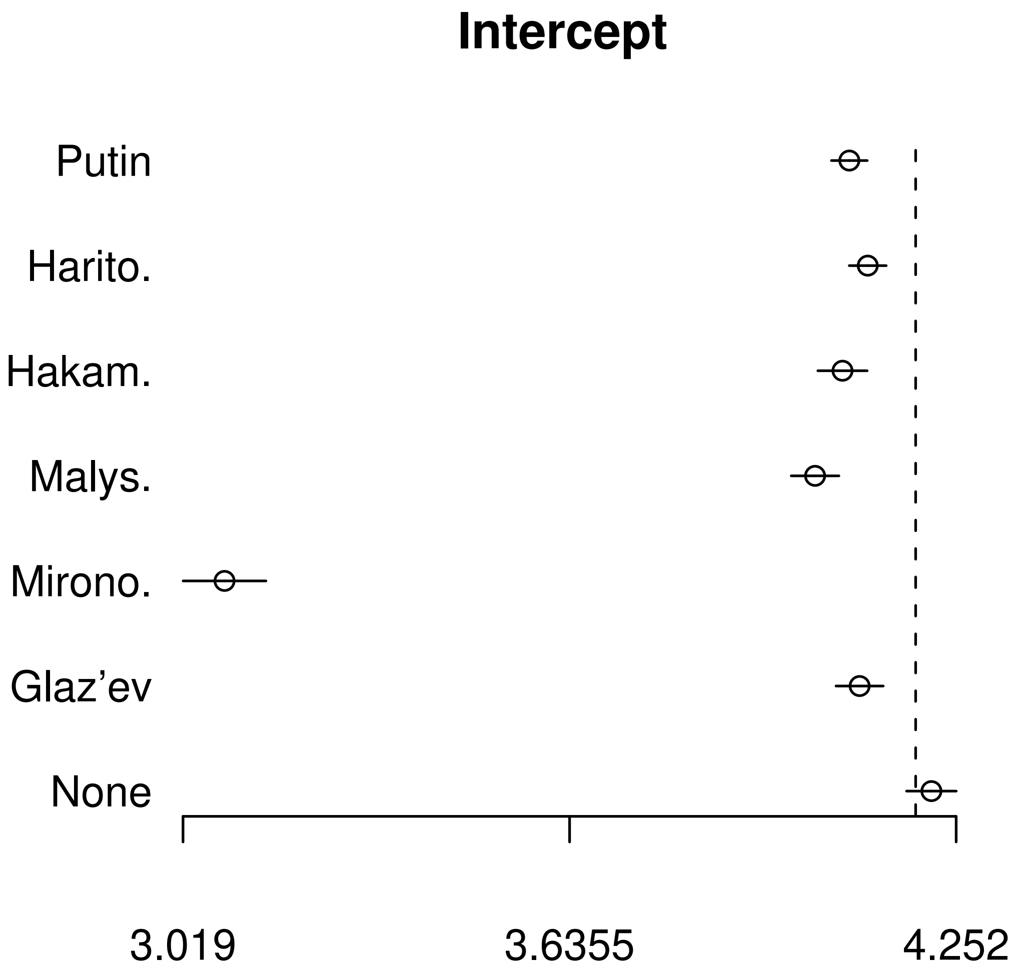

Figure 3: Second-digit means (robust weight model), 2004: intercept

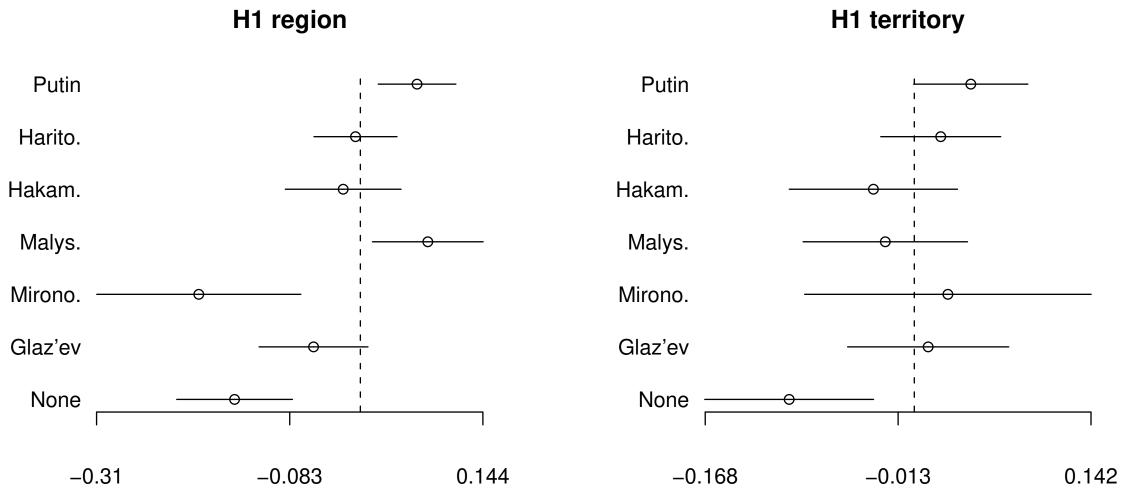

Figure 4: Second-digit coefficients (robust weight model), 2004: turnout

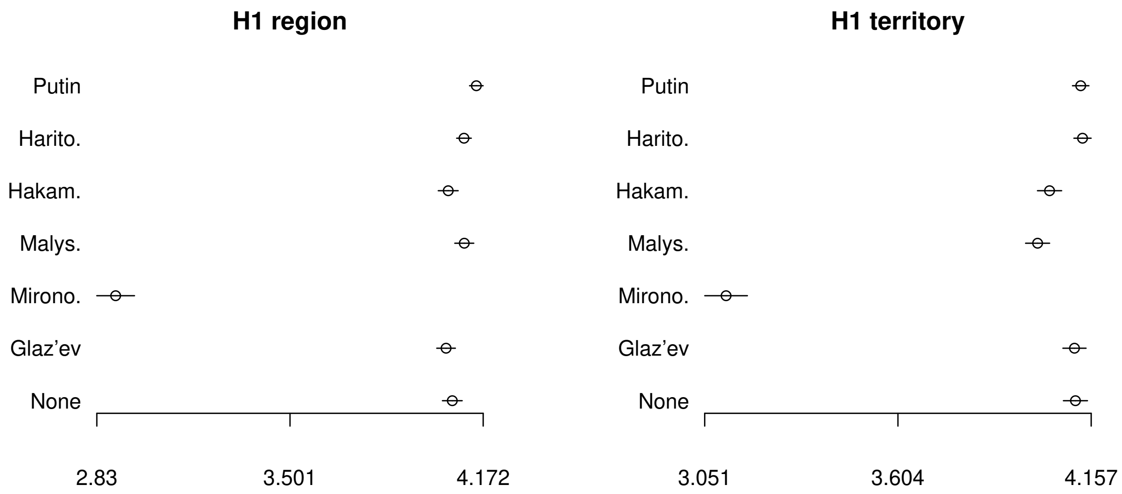

Figure 5: Second-digit means (robust weight model), 2004: turnout outliers

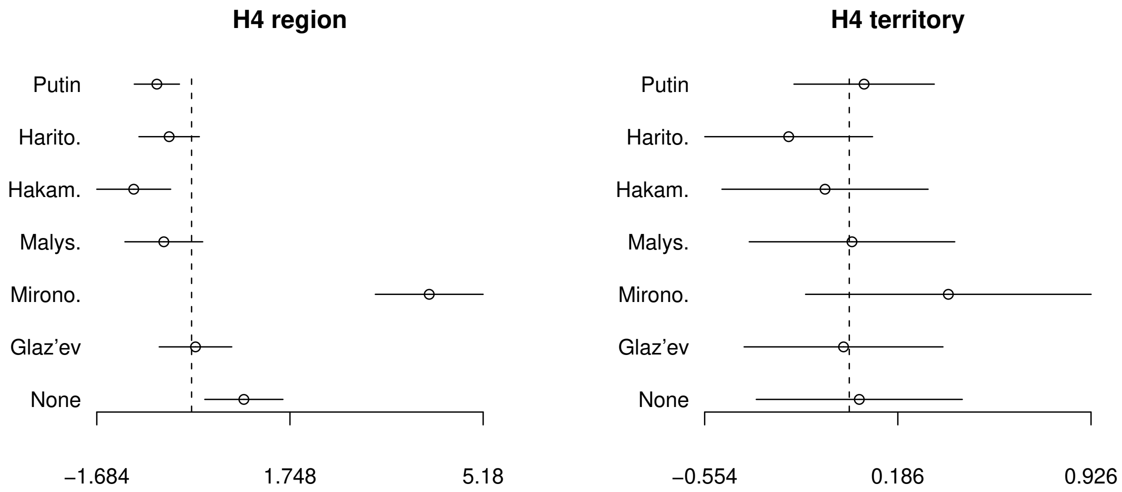

Figure 6: Second-digit coefficients (robust weight model), 2004: invalid/lost ballots

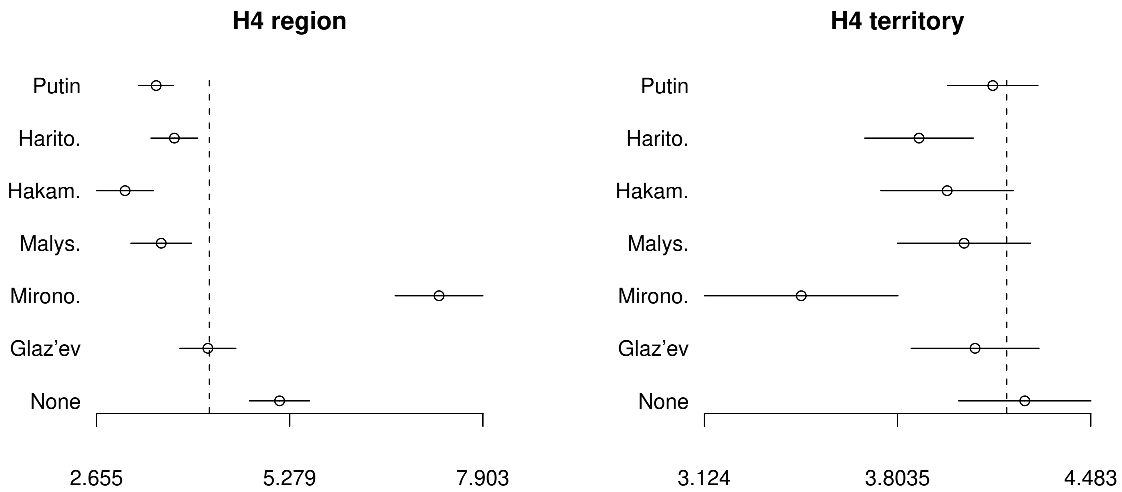

Figure 7: Second-digit means (robust weight model), 2004: invalid/lost ballot outliers

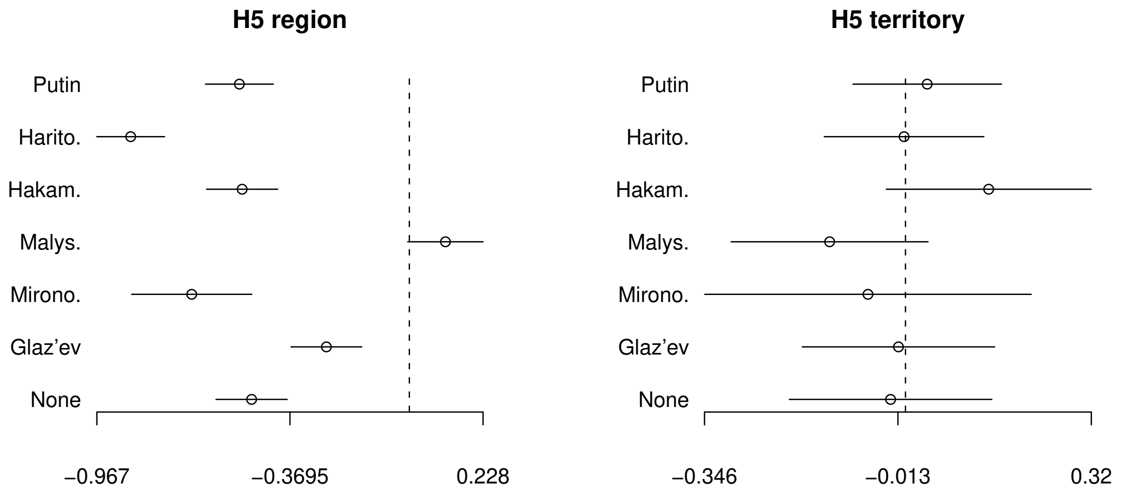

Figure 8: Second-digit coefficients (robust weight model), 2004: absentee certificates

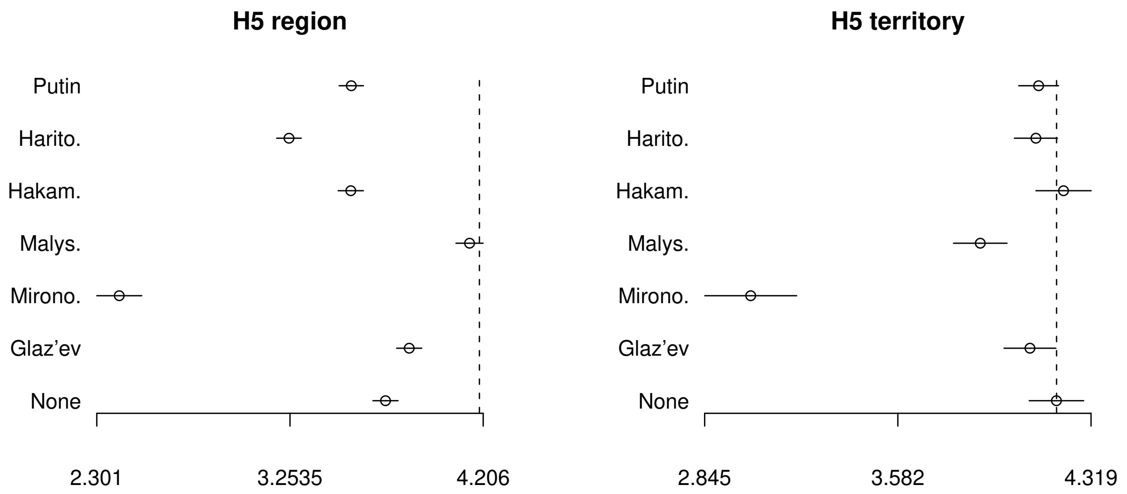

Figure 9: Second-digit means (robust weight model), 2004: absentee certificate outliers