Statistical fingerprints of electoral fraud?

Protests greeted the results of Russia's federal election in 2011, with accusations of vote-rigging and fraud. With the country heading to the polls again this September, **Dmitry Kobak**, **Sergey Shpilkin** and **Maxim S. Pshenichnikov** shine a light on several anomalies in the election data set

When Russians head to the polls in September this year to vote in elections for the State Duma, the lower house of the Russian Federal Assembly, there will be those left wondering whether the outcome of the vote can be trusted. At the last election, on 4 December 2011, protests erupted over claims that the elections had been rigged.

The United Russia party, led by the then prime minister – now president – Vladimir Putin, won 49% of all votes cast throughout the country. In Moscow, the capital, United Russia's share of votes was 47%. There were, however, numerous reports of electoral fraud – in Moscow in particular, with its large number of independent observers.

Protests in response to these claims culminated a week after the election, on 10 December, in a demonstration by an estimated 50 000 people – the largest the country has seen since the early 1990s. In the face of such protests, Russian officials denied any electoral manipulations, while numerous attempts to litigate the results, using observers' reports and smartphone-captured videos as evidence of fraud, were declined by courts on various grounds. In February 2012, for example, the Investigative Committee of the Russian Federation claimed that videos showing alleged acts of falsification of votes at polling stations "had signs of video editing", that is, they were themselves falsified.

With legal avenues closed, protests dwindled and life went on as before. But suspicions still remain – suspicions that have been further fuelled by statistical analysis.

Open *and* honest?

All electoral results in Russia are freely available online, down to the level of a single polling station: for each of some 95 000 polling stations, researchers can see how many people were registered to vote, how many came to vote, and how many voted for each party. This makes Russia a rare and curious case: a country where such information is open to all, and yet elections are often claimed to be manipulated. However, with all these data available, can we assess the fairness of elections by inspecting the raw election data?

Various researchers in Russia have attempted to do so since the early 1990s. One of the approaches, proposed by Sobyanin and Sukhovolsky,[1] is based on the following idea. Let us consider what happens with the practice of ballot stuffing – when a large number of fake ballots for some party (A) are added to the ballot box. In this scenario, the results of party A – defined as the ratio of the number of ballots in its favour to the number of all collected ballots – will increase. But so too will turnout – defined as the ratio of the number of participating voters to the number of registered voters.

Now, assuming that ballot stuffing of this kind happens only at a subset of polling stations, we should see a pronounced positive correlation between the result of party A and the turnout across all polling stations in the country. Sobyanin and Sukhovolsky suggested that the presence of such a correlation may be indicative of electoral fraud in favour of party A. Similar ideas have recently been followed up in modelling studies.[2]

Figure 1A shows turnout–result distributions for all seven major Russian elections (four presidential and three parliamentary) after the year 2000, when detailed data

Dmitry Kobak is a postdoctoral researcher in the Champalimaud Centre for the Unknown, Lisbon, Portugal. His research focuses on statistical analysis of neurophysiological recordings and on developing dimensionality reduction methods suitable for neural data

Sergey Shpilkin is a technical translator and writer in Moscow, Russia. Previously, he worked as a researcher specialising in quantum chemistry and quantitative structure–activity relationship modelling. Since 2007 he has also been an independent data journalist and election analyst

Maxim S. Pshenichnikov is employed by the University of Groningen, the Netherlands. His research focuses on a wide range of ultrafast phenomena in organic materials at nanoscopic lengths and femtosecond time scales. Statistical data analysis is one of his spare-time hobbies

Pshenichnikov photo © Sylvia Germes

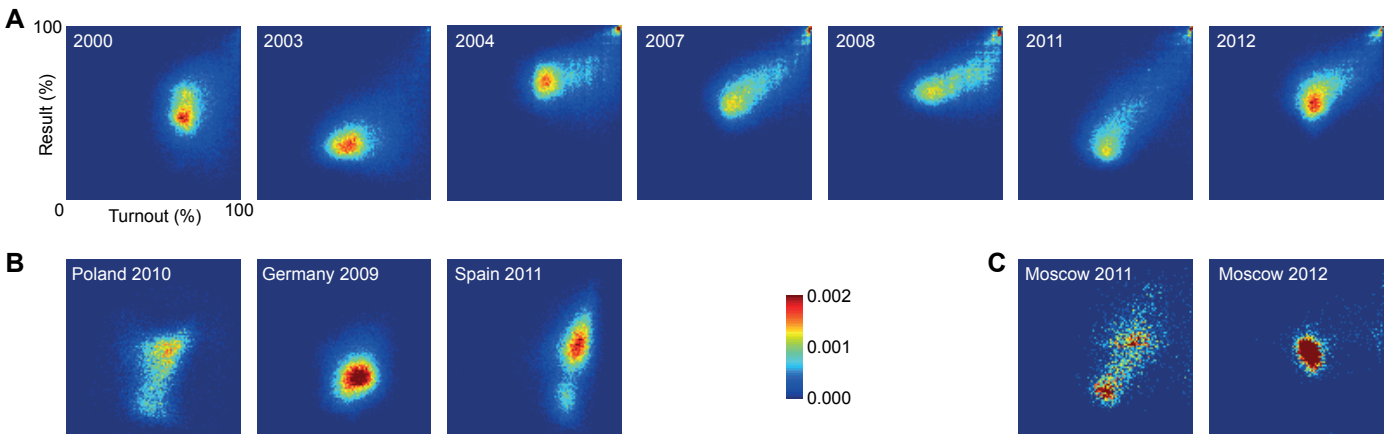

FIGURE 1 Turnout–result distributions. (A) Russian presidential (2000, 2004, 2008, 2012) and parliamentary (2003, 2007, 2011) elections. The horizontal axis on each plot corresponds to turnout, the vertical axis to leader's result. Each axis was split into 1% bins centred at integer percentages (0–0.5%, 0.5–1.5%, ..., 99.5–100%). Colour shows the number of polling stations in the corresponding 1×1% bin normalised by the total count, so that all numbers together sum to 1. Polling stations with 100% turnout were removed from the analysis because they include stations for which the number of registered voters was not available in advance (those were officially assigned 100% turnout). (B) The same for three European elections: the 2010 presidential election in Poland (first round), 2009 federal election in Germany (*Zweitstimmen*, i.e. party votes), and 2011 Congress election in Spain. (C) The same for the city of Moscow in the 2011 and 2012 Russian federal elections

became publicly available.[3] Turnout is on the horizontal axis on each plot, and leader's result is on the vertical axis; the "leader" is either Vladimir Putin or a party or candidate associated with and supported by him. Each axis is split into bins representing a single percentage point, and the 2D distribution of polling stations on this 100 × 100 grid is colour-coded. Strong correlation between turnout and leader's result can be seen starting from 2004, each case additionally enhanced by what looks like a separate cluster of polling stations with near-100% turnout and near-100% results in favour of Putin or his associates (creating a bimodal distribution).

However, the presence of correlation is not a proof of falsification *per se*, and neither is the bimodality of the 2D distribution. Indeed, Figure 1B shows three examples of European elections that have never been accused of falsifications: in Poland, Spain, and Germany (these countries were chosen because the number and size of polling stations is comparable to Russia, and the data are available with the same resolution). Noticeable correlation is present in all of them, and Spain exhibits a clear bimodality of the 2D histogram. These effects are likely due to geographical and socio-cultural inhomogeneities: it can happen that one part of the country has sufficiently different political attitudes than the rest, yielding correlations and/or bimodality. For example, the lower cluster in Spain comes from Catalonia and the Basque Country, and bimodality of votes in Poland is associated with the well-known division of the country into eastern and western parts with their different historical backgrounds.

Still, in some cases geographical inhomogeneity does not appear to be a plausible explanation. For example, very strong correlation and bimodality in the city of Moscow observed in the 2011 Russian elections (Figure 1C) vanished entirely in the following 2012 elections (held only 4 months later). As the socio-geographical situation in Moscow could hardly have changed in the time between these two elections, ballot manipulation might be considered a reasonable explanation of what gave rise to the 2011 distribution, and the "true" result of United Russia can be estimated to be around 30% (the position of the lower-left cluster in the 2D diagram), in good agreement with another analysis based on reports of independent observers.[4]

Correlation and bimodality are not definitive evidence of fraud, but there are other curiosities in the data worth exploring

Nonetheless, convincing as this argument might be to some, it is not definitive proof. Sceptics might rightly argue that voting patterns in the 2011 parliamentary elections are not necessarily comparable to those in the 2012 presidential elections. Indeed, a city that was split on the parliamentary choice could vote much more unanimously for the president.

A clearer picture?

If we accept that correlation and bimodality are not definitive evidence of fraud, there remain other curiosities in the data

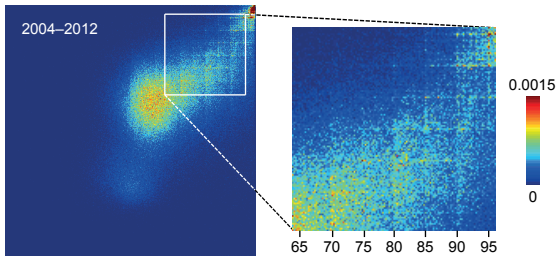

FIGURE 2 An average distribution over five elections from 2004 to 2012, calculated with a finer 0.25% bin size as compared to Figure 1

worth exploring. Looking closely at the two-dimensional distributions in Figure 1A, one might notice another, more subtle, peculiarity: seemingly periodic vertical and horizontal lines in the upper-right tails of the distributions, most apparent in 2008. This can more clearly be seen on the average histogram over all elections from 2004 to 2012 (Figure 2). In order to inspect this pattern more closely, we consider one-dimensional turnout and leader's result distributions (Figure 3) – horizontal and vertical projections of the two-dimensional histograms. Here the phenomenon becomes obvious: in all elections starting from 2004, both distributions exhibit pronounced peaks at multiple-of-five percentages such as 65%, 70%, 75%, and also smaller peaks at all sufficiently high integer percentages such as 91%, 92%, 93%.[3]

These peaks are observed at the same percentage values in several elections and are clearly periodic; this suggests that they are very unlikely to be the result of random fluctuations. We performed a significance test based on a Monte Carlo simulation of fair elections, and showed that the peaks are significant with $p<0.0001$.[3] In other words, there are specific percentage values of turnout and leader's result that are much more widespread than others, and these values happen to be "round": integer rather than fractional, and multiple of 5 or 10 rather than non-multiple.

There is a large amount of experimental and observational evidence showing that people are attracted to round numbers and, when choosing numbers, these are the ones they tend to go for. This is why, for example, the amount of money given in tips and the numeric answers to questionnaires tend to cluster near round numbers. Similarly, the prevalence of round percentages in the election results suggests that a substantial fraction of these results may have been consciously chosen instead of having arisen through the random voting process.

Imagine a polling station with 1755 registered voters (a typical value for a Russian city). Imagine further that election officials decided to forge the results at this particular polling

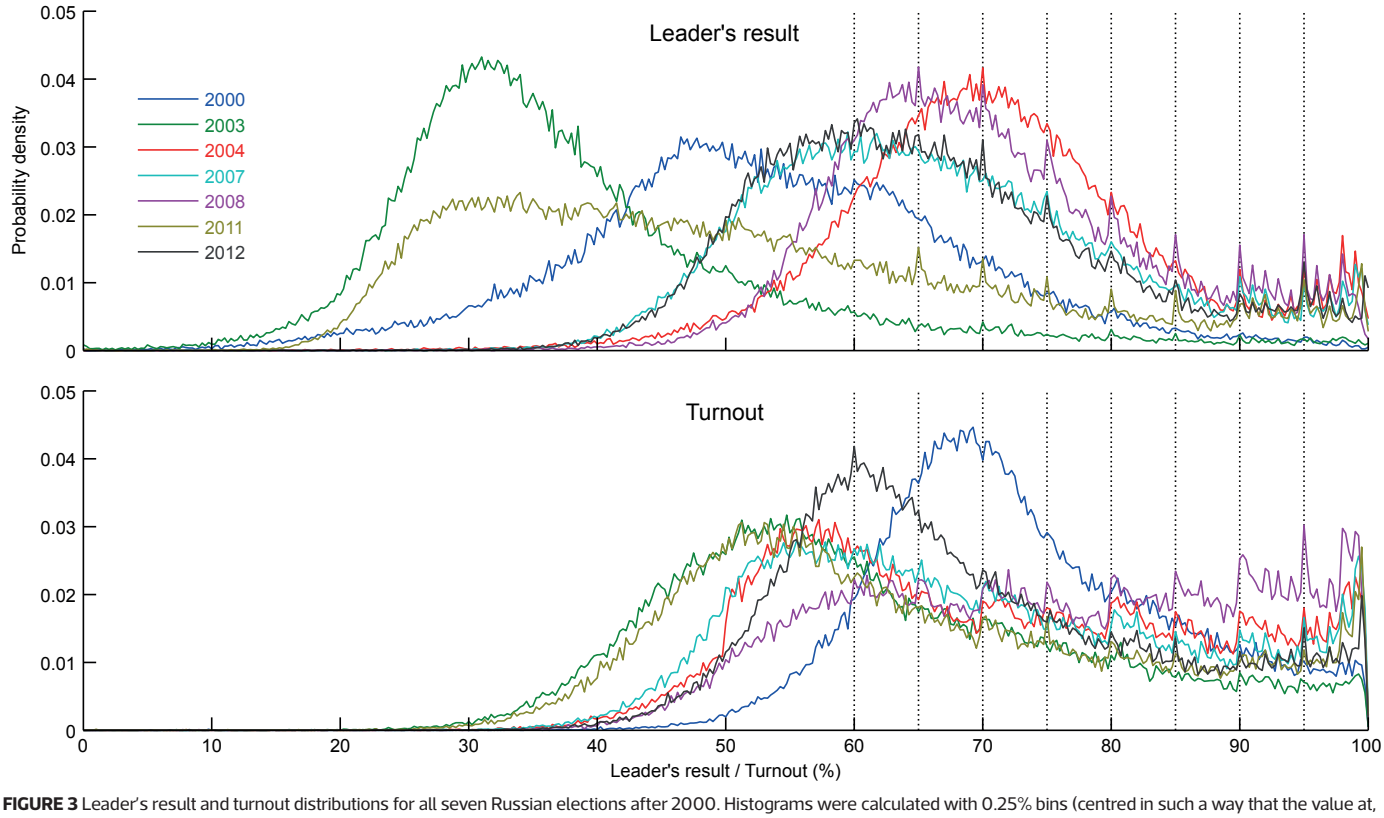

FIGURE 3 Leader's result and turnout distributions for all seven Russian elections after 2000. Histograms were calculated with 0.25% bins (centred in such a way that the value at, for example, 50% corresponds to 50 ± 0.125%) and normalised to have unit area. To avoid any artefacts due to integer division, we added a random number sampled from the uniform distribution $U(-0.5, 0.5)$ to the numerator of each turnout or leader's result fraction.[5] As before, polling stations with 100% turnout were excluded from this analysis

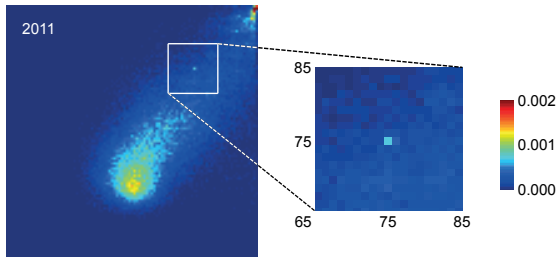

FIGURE 4 A zoom-in to the region around 75% turnout and 75% leader's result of the 2011 histogram from Figure 1A

station and report a turnout of 85%. They chose the number 85% because it is round and hence inherently appealing, and so other dishonest polling stations will be likely to use the same number. To achieve 85%, this polling station needs to report 1755 × 0.85 = 1492 ballots cast (as it is the number of ballots cast that is recorded in the election protocols, not the resulting percentage). Note that the number 1492 is not remarkable in itself; it is only the resulting percentage value (i.e. the 1492/1755 ratio) that is round.

It has previously been argued that distributions of individual digits in the ballot numbers across polling stations can reveal falsification fingerprints.[6,7] However, even though a large number of ballot counts ending in zero may be an indication that something is wrong, it does not necessarily imply purposeful falsification, as ballot counts could be rounded due to human error. In contrast, achieving round *percentage values* requires careful adjustment of the reported numbers. Thus, statistically significant prevalence of turnout and/or leader's results at "appealing" (integer or round) values may serve as strong statistical evidence for falsification.

The prevalence of round percentage values in the 2011 election results was noticed soon after the election; subsequently, a large part of the analysis presented in this article was carried out before the 2012 election. This means that the data from the 2012 election can be regarded as directly testing our *a priori* hypothesis. The unchanged presence of round and integer percentage peaks in the 2012 election data thus lends additional support to the theory.

Sometimes, the attraction to "good-looking" numbers reaches grotesque proportions. For instance, looking back at Figure 1A, one may notice a bright dot (a local peak in density) at the 75% turnout and 75% leader's result in the 2011 election (Figure 4). It turns out that this dot is contributed by a single city – that of Sterlitamak in the Republic of Bashkortostan, where 42 polling stations out of 107 reported 75 ± 0.5% turnout and leader's result. This observation suggests that the polling stations contributing to the round percentage peaks might be geographically localised, instead of being distributed all over the country. Indeed, we find that the prevalence of round percentages arises mostly due to the contribution of approximately 15 out of 83 Russian regions (the Republic of Bashkortostan being one of them), strongly suggesting that there is an orchestrated nature to these anomalies.[3]

An even more ludicrous example is provided by the city of Vladikavkaz in the same 2011 election. The city is divided into two constituencies. If we look at just one of these constituencies (Figure 5), we see that the vast majority of the polling stations reported 74% of votes for the United Russia party, while the Communist party obtained 20%. Curiously, at five polling stations this pattern was reversed, with the Communist party getting 74% and United Russia only 20%. This pattern seems highly artificial.

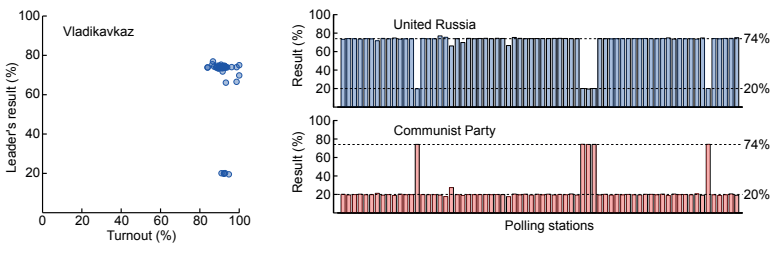

FIGURE 5 Election results in 2011 in one of the two constituencies forming the city of Vladikavkaz

Calling it

In summary, we have shown that statistical analysis of electoral data can uncover various irregularities that may indicate large-scale electoral manipulations. Most of these irregularities may have alternative explanations that can only be discarded on the basis of historical data, geography, and other information external to the statistical analysis. We argue, however, that the high prevalence of round percentage values in the election outcomes provides a statistical indication of fraud that cannot easily be explained away. The evidence is, in our view, beyond reasonable doubt.

Will these statistical irregularities show up in the upcoming federal election in September 2016? We shall see. ■

References

1. Sobyanin, A. A. and Sukhovolsky V. G. (1995) Demokratija, ogranichennaya falsifikatsijami: vybory i referendumy v Rossii 1991–1993. www.hrights.ru/text/sob

2. Klimek, P., Yegorov, Y., Hanel, R. and Thurner, S. (2012) Statistical detection of systematic election irregularities. *Proceedings of the National Academy of Sciences of the USA*, **109**(41), 16469–16473.

3. Kobak, D., Shpilkin S. and Pshenichnikov M. (2016) Integer percentages as electoral falsification fingerprints. *Annals of Applied Statistics*, **10**(1), 54–73.

4. Enikolopov, R., Korovkin, V., Petrova, M., Sonin, K. and Zakharov, A. (2013) Field experiment estimate of electoral fraud in Russian parliamentary elections. *Proceedings of the National Academy of Sciences of the USA*, **110**(2), 448–452.

5. Johnston, R. G., Schroder, S. D. and Mallawaaratchy, A. R. (1995) Statistical artifacts in the ratio of discrete quantities. *American Statistician*, **49**, 285–291.

6. Mebane, W. R. (2006) Election forensics: Vote counts and Benford's law. Paper presented to the Summer Meeting of the Political Methodology Society, UC-Davis.

7. Beber, B. and Scacco, A. (2012) What the numbers say: A digit-based test for election fraud. *Political Analysis*, **20**, 211–234.