The University of Chicago

# Putin Ballots in the Box: A Geospatial Analysis of Manipulation in Russian Elections

by
Stasya Rodionova
srodionova@uchicago.edu

April 22, 2022

*A thesis submitted in partial fulfillment of the requirements for the degree of*

Bachelor of Arts in Political Science at the University of Chicago

**Preceptor: Professor Patricia Conley**
**Advisor: Professor Scott Gehlbach**

**Abstract**

Vladimir Putin's twenty year long regime has time and time again faced scrutiny and criticism from international election observers and citizens alike. Countless videos of ballot stuffing have emerged on the internet and many suspect that the election results reported by the Russian government are fraudulent. Building on previous election forensics projects studying Russian elections, this thesis aims to expand understanding of the geographic spread of electoral fraud in Russia. Using a stochastic kernel density resampling method, I estimate regional levels of contamination in Russian presidential elections from 2000-2018. The indicator of interest is round integer (multiple of 5) reports of voter turnout and incumbent vote share. Mapping the relative frequency of round integer reports of vote share across the election years, I determine that Russia's ethnic republics most consistently report abnormally high turnout and vote share. I find that the amount of election manipulation has increased throughout the course of Putin's regime. Moreover, statistical analysis points to the bifurcation of the electorate, signifying that regions which had previously reported abnormal results are doing so at much higher rates in subsequent election years. This is reflected in the Kernel Density Estimates for each year, which demonstrate abnormal spikes at round integers and show a remarkable shift towards a non-normal distribution over the years of Putin's regime.

**<u>Table of Contents</u>**

**Introduction and Problem Statement**

Russia's transition towards an electoral democracy has been, put bluntly, unsuccessful by a lot of metrics. Both Freedom House and The Economist's Democracy Index have categorized Russia as "not free" and "authoritarian," respectively. This is reflective of a downward trend that the country has been experiencing since the early 2000s as recorded by global democracy and freedom indices.[1] Personal and democratic freedoms have been on the decline in a country that continues to host elections despite the fact that one man and one party have remained in power since his election nearly a quarter of a century ago.

Over the years OSCE and other independent observers have cautioned that Russian elections have been deficient in meeting the standards of free and fair elections. With every election, evidence of ballot stuffing, bussing voters to polls, and other forms of voter intimidation and falsification emerges and is recorded. Yet quantifying the amount of fraud is an elusive endeavor. Despite how interesting it may be to figure out to what degree Putin's (and United Russia's, his preferred party) popularity is real, working backwards from falsified elections data is an unreliable way to evaluate the extent of election falsification present in Russian elections.

Even though we may believe we understand why autocrats seek to legitimize their regimes via elections and we can describe how they ensure that their rule remains unchallenged, finding evidence of this behavior in the real world can sometimes be rather challenging. Acts of executive aggrandizement when it comes to acquiring institutional power are relatively easy to spot (the dissolution of a parliament, replacement of the judiciary), but the undermining of citizens' choices via secret ballot is much harder to detect. In fact, even exit polling in Russia has

---

[1] Ivanov, Mark. "Dynamics of Russia's Position in Democracy and Freedom Indices: Research and Approaches to Explanation." *JOURNAL OF GOVERNANCE AND POLITICS*, 9 Dec. 2021, https://sgpjournal.mgimo.ru/2018/2018-2/dynamics-of-russias-position-in-democracy-and-freedom-indices

become notoriously unreliable. Rather than asking citizens about their opinion on the issues or how they voted, exit polls ask voters, for example, to rank proposed constitutional amendments by "importance" to them, whatever that may mean.[2] This sort of data manipulation and cherry picking of indicators does little to target the real questions of consequence, which would allow us to compare Putin's hypothetical popularity according to the vote percentages he wins with the reality of how he is perceived. Thus, measuring the legitimacy of the vote percentages that Putin and United Russia win through public opinion polling is unreliable. These problems are common to many other hybrid regimes.

Competitive authoritarian systems are incredibly stable.[3] Part of the reason why is their continued usage of elections as a source of legitimacy. Evidence of ballot stuffing and other offenses is useful but estimating the scale of the fraud based on several recorded cases of misconduct is challenging. Finding evidence of this sort of behavior in the statistical fingerprint left behind by those that alter the numbers in favor of the incumbent regime, however, is something that would allow us to further explore how democratic autocracies go about this process and whether the way that falsification is carried out has evolved in recent years. At the center of this is a question of whether democratic autocracies adapt to sustain themselves and, if they do, whether we can chart that in the trails they leave behind.

The question I aim to answer is how has the geography of election fraud in Russia changed throughout the course of Putin's regime? To what extent can we say that the places and

[2] Yudin, Greg. "In Russia, Opinion Polls Are Being Used to Cover up a Divided Society." *OpenDemocracy*, 2 July 2020, www.opendemocracy.net/en/odr/russia-opinion-polls-referendum/.
[3] "Elections, Protest, and Regime Dynamics." *Elections, Protest, and Authoritarian Regime Stability: Russia 2008–2020*, by Regina Smyth, Cambridge University Press, Cambridge, 2020, pp. 1–24.

practices of election manipulation in Russia have changed throughout the past 20 years? Are there any spatial or temporal trends in election manipulation prevalence in Russia?

The main drive of this question is to take a closer look at how election falsification behaves spatially in Russian elections. I look for regional clusterization of fraud and whether the clusters, if they exist, have shifted throughout the past 20 or so years of Putin in power. It is possible that election manipulation has moved across regions. As the population of certain cities and regions fluctuates and Russia trends towards further urbanization, voting patterns may have changed as well, and with that, the way that fraud presents itself in Russian elections.

My expectation is that election fraud in the early 2000s was mostly concentrated in what Russians call "the periphery" (i.e. remote, rural, sparsely populated regions of Russia) but that in more recent years evident manipulation has spread closer to urban areas. The potential clusterization of election fraud in remote regions makes intuitive sense because these are areas out of the way and therefore possibly less accounted for by independent elections observers. They also may have additional incentive to demonstrate loyalty to Moscow in exchange for potential benefits in resource allocation. Cities, on the other hand, are more likely to be under careful watch by third parties and thus present more of a challenge.

The reasons behind this shift of fraud from remote to more populated areas are numerous. One potential explanation is that election fraud in a particular region could "infect" the rate of election fraud in neighboring voting districts. This project, however, does not aim to answer what is responsible for regional clusterization or its movement. The goal is simply to determine whether there has been any change in the way that election fraud presents itself in Russian elections regionally over the past ~20 years.

**Literature Review**

*Detecting Electoral Fraud in Russia*

One of the most well-known researchers of Russian election fraud is independent researcher Sergey Shpilkin, a physicist that does statistical analysis of Russian election results. His research analyzes "peaks" in election results (in vote shares, turnout percentages, etc.).[4] He builds histograms of percentage of votes cast for Putin vs total votes cast at particular polling stations and looks for abnormalities in them. Shpilkin finds that there are spikes at 65% turnout and 62% for the leader's result in the elections. These spikes constitute what is known as Churov's Saw and they appear at intervals that are too regular to be naturally occurring in a data set of this size.

These numbers not only present statistical abnormalities because they are too "neat" for real world election results, but also because they tend to correlate by region and present a whole host of other red flags. What these abnormalities show us is that there are likely certain quotas that need to be met in election results and turnout that are passed down through the rungs of the Russian government, resulting in the spikes we see in the data. Round numbers are a rarity in these sorts of samples, so turnout or vote percentages in favor of a particular candidate at non-fractional numbers are suspect. Furthermore, high turnout rates in districts that have low political engagement are another concerning factor, especially in remote areas where access to polling stations may be difficult. Specifically, ethnic republics (some of which are remote mountainous regions or embroiled in conflict) report an overwhelming turnout rate that cannot be rationally explained. These results demonstrate foul play rather than reflecting the genuine

[4] Kobak, Dmitry, Sergey Shpilkin, and Maxim S. Pshenichnikov. 'Putin's Peaks: Russian Election Data Revisited'. *Significance*15, no. 3 (2018): 8–9. https://doi.org/10.1111/j.1740-9713.2018.01141.x.

interests of the population. Shpilkin's research shows us a method of flagging election fraud because ballot stuffing leaves behind a distinctive pattern, one that is characterized by statistically improbable round numbers and turnout and vote share ratios.

Shpilkin has replicated this analysis for both presidential and duma elections throughout Putin's regime and found more of the same irregularities. His 2016 paper with Kobak and Pshenichnikov finds that there is a distinct "fingerprint" of fraud that can be detected in the election results.[5] This fingerprint is an unusual bimodal distribution in both turnout and leader's result. When the authors compared Russian election results to election results from other countries, they found that the Russian plot of turnout versus vote share demonstrated a distinct grid line at the intersections of round numbers and multiples of 5, something that was absent in the election results of other countries. That is, that there is a significant density of polling stations that are reporting integer values for both turnout and vote share, resulting in the visualization having distinct repeat x and y coordinates (turnout and vote share) at integer values rather than being naturally distributed. They also find that the 2011 elections in Moscow (a year of protests) have a distinct bimodality which cannot be explained by geographic inhomogeneity because the same phenomenon does not appear in the 2012 elections held in Moscow, which were held only four months later. The issue with bimodality is that it is not reflective of the typical electorate. Most election results are (and can be expected to be) normally distributed in terms of turnout rates and voting preferences. One notable exception to that is Canadian elections. However, this is explained by the fact that French-speaking Canada (Quebec) and the other provinces are different enough in their preferences and political habits that the country's electorate begins to look non-homogenous. The same cannot be said for Russia, however. With hundreds of ethnic

[5] Dmitry Kobak, Sergey Shpilkin, and Maxim S. Pshenichnikov. 2016. "Statistical fingerprints of electoral fraud?" *Significance* 13 (4), doi: 10.1111/j.1740-9713.2016.00936.x.

groups and languages spoken within Russia, there is no clear natural way for the electorate to split into two, which is why we should be suspicious of bimodality in Russian elections.

In addition to this finding, Shpilkin, Kobak, and Pshenichnikov report that the probability density charts of turnout and vote share have increased in "spikiness" (an indicator of fraud) since 2000 and 2004. Visually, this is evident through an increase in the number of individual spikes over the years at round and whole integers for vote share and turnout rates. Moreover, the increase in spikiness in Russian elections is not randomly distributed. Rundlett and Svolik carried out an analysis of Russian election results and found support for their claim that falsification in Russian elections is demonstrated partially through an abnormal volume of precincts reporting results that are multiples of 5. They hypothesize that this is due to human propensity for round numbers.[6] They demonstrate that the unusual volume of precincts reporting these round numbers is highly statistically improbable and cannot be attributed to chance. Not only are the spikes indicative of election manipulation becoming more frequent over the years, they are also occurring at points that seem way too good (and round) to be true.

*Geopolitical Background of Russia*

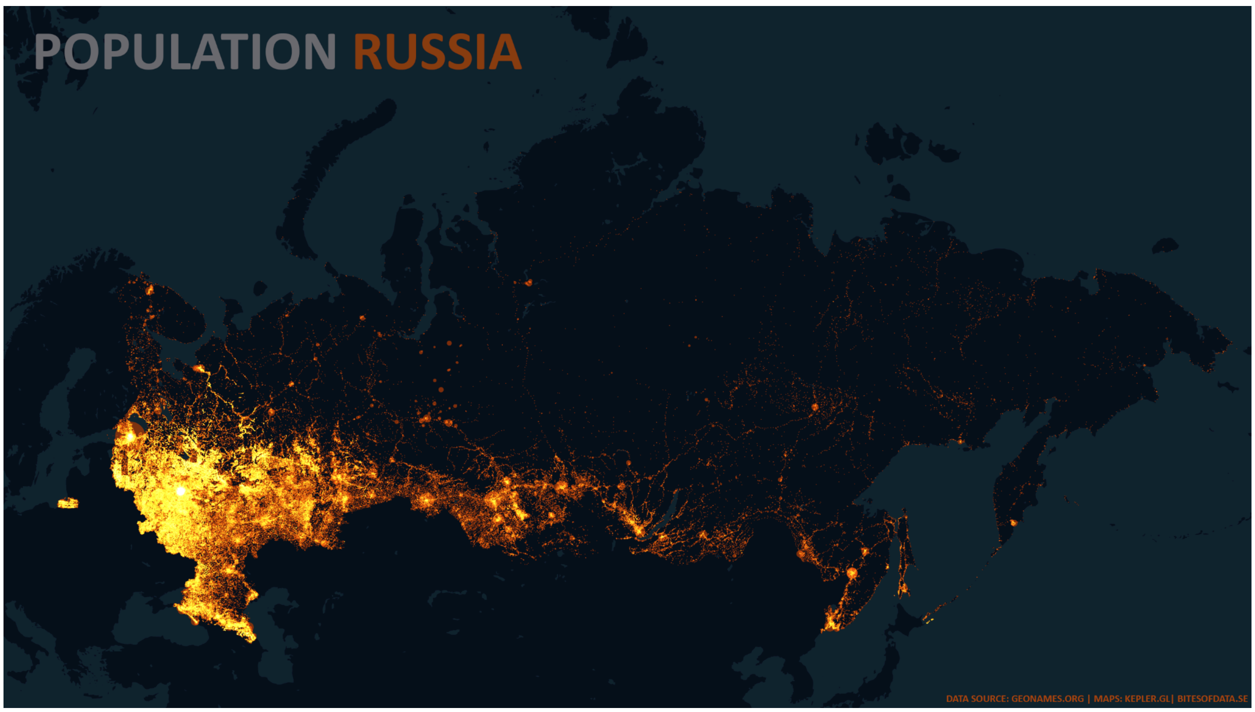
To understand the behavior of election fraud in Russia and the way the electoral system functions it is important to understand Russian ethnic, geographic, and administrative divisions first. Russia is broken into approximately 83 administrative regions, although this number has fluctuated in the past. These divisions are known as federal subjects and for the American audience can best be understood as the equivalent of states when it comes to elections. These federal subjects are further classified into oblasts, krais, autonomous okrugs, republics,

[6] Rundlett, Ashlea, and Milan W. Svolik. 'Deliver the Vote! Micromotives and Macrobehavior in Electoral Fraud'. *American Political Science Review* 110, no. 1 (February 2016): 180–97. https://doi.org/10.1017/S0003055415000635.

autonomous cities, and autonomous oblasts. For the purposes of elections as well as for the purposes of this paper, all of these regions are essentially legally equivalent. The borders of the federal subjects are not necessarily defined by population size, but rather are historic boundaries remaining from the Soviet Union's administrative divisions. Some of these federal subjects, however, have distinguishing features. For example, special status has been granted to the cities of Moscow and St. Petersburg, which are their own individual federal subjects. An 84th federal subject is the aggregate district of all overseas voters (which has been dropped from the data set for the purposes of this project). Additionally, the city of Baikonur in Kazakhstan participates in Russian federal elections for historic reasons and has also been dropped from the dataset. Finally, the city of Sevastopol and the Republic of Crimea are not internationally recognized as legitimate Russian subjects and so are also removed from this dataset. The map below shows the breakdown of the administrative federal subjects.[7] [8]

[7] *Administrative-Territorial Structure of the Russian Federation*. Iris Press, 2016.
[8] N.B. This map includes the Republic of Crimea and the city of Sevastopol. Russia's claim over these territories is not internationally recognized.


The next most important geopolitical facts to know about Russia concern its population density and ethnic makeup. The geographic expanse of Russian territory spans two continents and is divided into two parts by the Ural Mountains. Because there are large swaths of Russia that are largely uninhabitable for humans due to the harsh climate, the majority of people live in the European part of the country, concentrated mostly around Moscow and St Petersburg. The Eastern part of the country is populated mostly in the south. A stretch of cities including Omsk, Novosibirsk, Krasnoyarsk, Chita, Khabarovsk, and Vladivostok lie in the south connected by the Trans Siberian Railroad. Another significant portion of the population lives in the South around Sochi and the Black Sea. To contextualize the distribution of the population further, it should be noted just how disproportionately large the populations of Moscow and St Petersburg are compared to the rest of the country. Moscow's population is roughly 12 million people and St Petersburg has approximately 5 million inhabitants. The third largest city, Novosibirsk, hovers at around 1.6 million inhabitants. Most other major cities have roughly 1 million inhabitants.[9] Roughly 75% of the Russian population lives in urban areas.[10] The map below illustrates the population density of the country.[11]

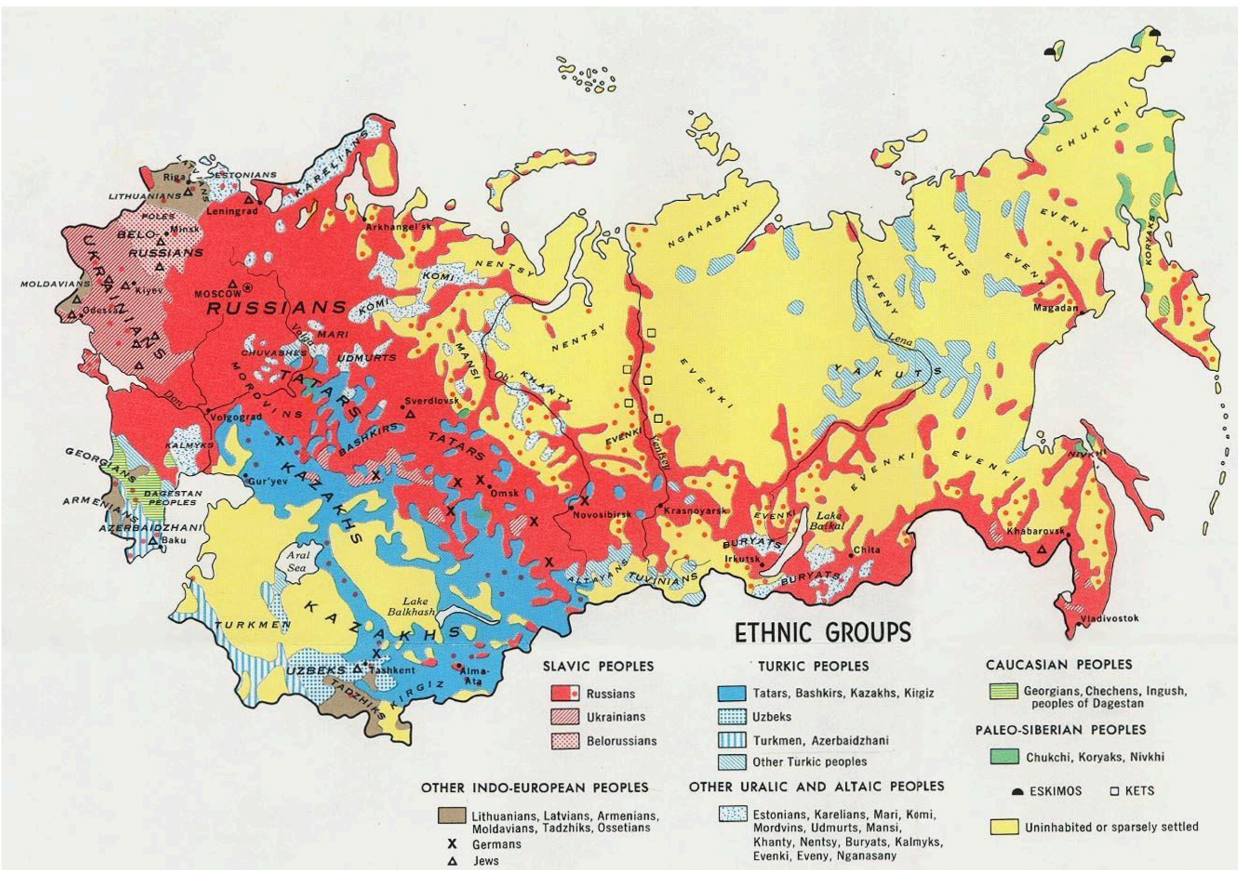



Additionally some of the federal subjects of Russia are roughly aligned with the distribution of ethnic minorities. The major ethnic groups in Russia are Russians, Tatars,

---

[9] "Russia: Largest Cities as of 2021." *Statista*, 11 Aug. 2021, https://www.statista.com/statistics/1090061/largest-cities-in-russia/.
[10] O'Neill, Aaron. "Russia - Urbanization 2020." *Statista*, 1 Feb. 2022, https://www.statista.com/statistics/271343/urbanization-in-russia/.
[11] Alex. "Population Density of Russia." *Vivid Maps*, 22 Feb. 2022, https://vividmaps.com/population-density-of-russia/.

Ukrainians, Bashkirs, Chuvashs, and Chechens, in that order. Although nearly 81% of the population self-identifies as ethnic Russians, the country is home to over 100 ethnic groups and has 35 official languages.[12] The 21 administrative republics of Russia are meant to correspond to specific ethnic groups. For example the Republic of Bashkortostan is home to the Bashkir people. However, many of these ethnic republics are no longer majority minority ethnic groups due to urbanization and migration patterns. Russia also has 5 autonomous okrugs which have a significant ethnic minority population. The map below shows the distribution of major ethnic groups of the Soviet Union.[13]




---

[12] "Russia Population 2022." *Russia Population 2022 (Demographics, Maps, Graphs)*, https://worldpopulationreview.com/countries/russia-population.

[13] Imgur. "Ethnic Groups in Russia [1342x941]." *Imgur*.

Notably, the Republic of Chechnya has had a complicated history with Russia. In 1994 the region fought for its independence from Russia in the First Chechen War and then again from 1999-2009 in the Second Chechen War, a key part of Putin's political platform in the early years of his presidency. This is especially relevant to the study of election manipulation in Russia given that Chechnya has historically reported record high turnout and record high Kremlin support. The truth of these numbers are dubious at best and laughable at worst given that Chechnya had for long been embroiled in a fight for its independence and was literally invaded and occupied by the Russian military, casting doubt on voter turnout rates of 99% and support for Putin and United Russia at 99%.[14] This and other interesting trends in the vote share and turnout reporting of other Republics and federal subjects of Russia are the premise for this project.

*Geographic Behavior of Fraud in Russia*

The basis for research into election forensics in Russia was greatly advanced by Ordeshook, Myagkov, and Shakin, who released a comprehensive guide to election forensics based on the results of their analysis of Russian elections.[15] In their explorations they not only find that patterns evident in the Russian vote are highly statistically improbable, but also find evidence for the increase of election manipulation from the year 2000 to 2004. Furthermore, they find that fraud is not committed uniformly across the country. They plot the distribution density of reported vote share and turnout to find that ethnic republics disproportionately committed fraud in 2000 compared to oblasts. In the year 2004 they found that oblasts had also begun

[14] Keating, Joshua. "Chechen Precinct Gives 107 Percent." *Foreign Policy*, Foreign Policy, 6 Mar. 2012, https://foreignpolicy.com/2012/03/06/chechen-precinct-gives-107-percent/.
[15] Myagkov, Mikhail, Peter Ordeshook, and Dimitri Shakin. *The Forensics of Election Fraud: Russia and Ukraine*. Cambridge University Press, 2009.

manipulating their results. This is demonstrated through a clear bimodality in the election results data, allowing them to point out that the electorate is bifurcating. In other words, the electorate is becoming more extreme in reporting high percentages of turnout and vote share in one sub-group of precincts and low turnout and comparatively low incumbent vote share in others. The electorate has two different groups (clusters) headed in two different directions in their reported turnout and incumbent vote share.

One of the most important takeaways from Ordeshook et al. is the distinction between ethnic republics and oblasts in the levels of fraud detected. Beyond the fact that ethnic republics overwhelmingly were subject to more manipulation, the authors show that in 2004 manipulation increased in rural parts of oblasts. In other words, they report a trend that seems to relate the level of fraud with distance to the nearest urban area or regional center in an oblast. They argue that this could be a direct impact of the influence of regional bosses over the urban economy. Historically, republics are majority minority ethnic groups. They tend to be in more remote regions and thus rely on resource redistribution by the center. There could be an added incentive to manipulate elections in republics, driven not only by the fact that it is unlikely that anyone will raise a fuss about the election results of some of Russia's most remote regions, but also by the hope that by supporting the incumbent and demonstrating loyalty, the republics (or their regional leadership) can benefit from more advantageous resource distribution.

Building on these findings, Mebane and Kalinin establish a method that helps map where the fraud is occurring.[16] They propose a spatial scan statistic which measures whether what is found in a given geographic window is significantly different from what is found immediately

---

[16] Mebane, Walter R. and Kalinin, Kirill, Geography in Election Forensics (2014). APSA 2014 Annual Meeting Paper, Available at SSRN: https://ssrn.com/abstract=2452260

outside of it. For this they use Monte Carlo simulations and geographic weights based on Euclidean distance to the nearest cities. They classify the Russian election data into binary categories (1 = if vote share for Putin ended in a 0 or 5, 0 = if it did not). They map it by region and find that there are 10 identified clusters in 2012, with 3 of them passing Monte Carlo significance (Dagestan & Chechnya, St Petersburg, Bashkortastan). They attribute this to the effectiveness of regional political machines in the republics. They emphasize the need to include geographic weights in election analysis and recommend Monte Carlo simulations but are hesitant about the practicality of the application to more complicated datasets.

Interestingly, Kobak, Shpilkin, and Pshenichnikov ran a similar test in 2016 and find different clusters. They use the frequency of reported round percentages in voter turnout and percent vote share as their measure of fraud. They then confirm that the “peaks” they see at unnaturally round numbers are statistically improbable by running Monte Carlo simulations. They find that the peaks repeat for the same exact integer locations in elections after 2004. They find that the top 15 most anomalous regions contribute the most to these repeat irregularities and that without them the pattern is significantly reduced. The top 15 regions were centered around Moscow, the south-western border with Ukraine, and the Caucasus. This is unlike the clusterization that Mebane and Kalinin reported in their analysis in 2014. Part of the reason why this is so can likely be attributed to the spatial weights that Mebane and Kalinin used.

In any case, study of the geographic distribution of election manipulation in Russia has been elusive. This project will aim to fill some of those knowledge gaps and provide one potential method of measuring relative prevalence of election manipulation across Russia’s regions. This project will generate the first spatial visualization of where electoral fraud in Russia is most present and how it has shifted from year to year. Following the methods used by

Rundlett, Svolik, and Rozenas, this project produces a first of its kind visualization of fraud for presidential elections since the year 2000.

**The 2008 Presidential Election**

It should be noted that this project also considers the vote share results from the year 2008. Although Vladimir Putin was not a candidate in the 2008 presidential elections, many consider his frontrunner and former Prime Minister, Dmitry Medvedev, to be a proxy for Putin, who had stepped down from his position several months prior. It is widely understood that Putin stepped down from his post as president of Russia in order to be eligible to run for president in the future. Putin was constitutionally barred from serving a third term, so his taking up the post of prime minister in 2008 to Dmitry Medvedev is considered a tandemocracy. This joint rule arrangement did not come as a surprise to the electorate, as Medvedev announced that if elected he would appoint Putin as his prime minister. Moreover, as the Levada Center found in a survey from 2008, most Russians believed that it was Putin that actually held power during the Medvedev presidency.[17] For these reasons, this project will consider the presidential election of 2008 to be an extension of the electoral mechanisms and patronage system established under Putin's first two terms even though it was not necessarily Putin who was on the ballot in that election.

[17] "Levada Center: Presidential Performance 2008." *Russia Votes*, https://www.russiavotes.org//president/presidency_performance.php#540.

## Methodology & Data

### *The Data*

Presidential election data was sourced from Dmitry Kobak's GitHub repository, which contains all Russian election data from the years 2000-2020.[18] This data was scraped from the website of the Central Election Commission of Russia by Sergey Shpilkin. To my knowledge, this is the most complete existing dataset containing Russian election results. The data contains the following relevant columns for each data point: region, tik (district ID), uik (precinct ID), total number of registered voters, total ballots received, number of valid ballots, number of invalid ballots, votes cast for Putin (or in 2008 Medvedev), and votes cast for all other candidates (listed out for each one individually). I generate the following columns: turnout *((valid ballots + invalid ballots) /total registered voters) *100)*, incumbent vote share (*(votes cast for Putin/valid ballots)*100*).

### *Measures of Electoral Manipulation*

The primary methods of election falsification that are used in Russian elections and can be detected with statistical tests are falsification of voter turnout and incumbent vote share. The method I use is an adaptation of methods developed by Rundlett and Svolik. The authors kindly provided a copy of their final Stata replication code, part of which I utilize for the purposes of this thesis.

Rundlett and Svolik first round each reported vote share and turnout data point to the nearest multiple of 0.5 using bins of size ± 0.05 and extract the single digit and decimal unit for each case. For example, 76.481 and 46.532 both become 6.5. After constructing a distribution of

[18] https://github.com/dkobak/elections

these decimal points (0.0-9.5), they find that there is an overrepresentation of precincts reporting vote share that lands on multiples of 5 (specifically at the points 0.0 and 5.0). They then use a Kernel Density Estimate to measure the difference between the actual reported frequencies of precincts reporting round digit vote share and the expected distribution. This difference is called the "ruggedness" of the distribution of Putin's precinct-level results. The authors drew their conclusions based on the ruggedness of the Kernel Density Estimates.

In this paper, I collapse raw election results into bins size ± 0.05 for convenience. I drop all precincts reporting 100% turnout (Russian voting precincts sometimes include prisons or hospitals, which for complicated reasons report 100% turnout and should be left out of the analysis). Additionally, I drop regions outside of the Russian Federation, including overseas voters, Baikonur, and Sevastopol.

For each election year I generate a cleaned-up version of the original csv file in Stata. The variables of interest are the region, tik ID (district), uik ID (precinct), percent turnout (generated by *((validballots + invalidballots) /totalregvoters) *100*), and Putin vote share (*(Putinvotes/validballots)*100*).[19] From there I generate rounded turnout and vote share numbers using bins size ± 0.05. I then generated a variable for the unique number of tiks in each region. I create a variable (v4) that divides the reported vote share of each precinct by 5 and replaces the values with "1" if it was a multiple of 5 and "0" if it was not a multiple of 5. I then find the percentage of precincts in each region that reported results that were multiples of 5 (mean of v4).

By taking the number of precincts with the specific abnormality of reporting turnout and vote share as a multiple of 5, we can get an approximate estimation of round-digit reporting fraud in each region. This is not to say that fraud is not committed by reporting falsified numbers

[19] Medvedev as proxy in 2008.

that aren't multiples of 5. After all, many ethnic republics report 98% turnout and vote share. Rather, this is used as a tool to narrow down the trends that I am looking for and to make the analysis viable for replication across all 85 of Russia's regions and all 5 election years selected. Using this proportion I construct choropleth maps for presidential election years 2000, 2004, 2008, 2012, and 2018 which map the relative density of round digit reporting precincts.

I then run robustness checks on the data. In a sample size this large, we expect there to be some naturally occurring amount of multiples of 5 reported for turnout and vote share. To account for this, I use a Kernel density resampling method to calculate the likelihood of multiple of 5 results and compare them to the data set. For this I use an R package written by Arturas Rozenas. This yields RKD-based estimates of fraud. The model works by comparing the kernel density estimates drawn from the observed vote share reports against the kernel density of a sample with null distribution. Building upon a stochastic model of elections, the result is an estimated percentage of contamination for each year or region run. This method applies specifically to vote share reporting.[20] The method developed by Rozenas is one of the more conservative methods of estimating election fraud. While many other methods will report higher numbers of election contamination, this one errs on the side of caution and thus presents a cautious estimate of fraud (result contamination) in Russia.

Mapping the frequency of the estimated contamination will show where the fraud is occurring and at what magnitude it is being carried out. Furthermore, this project adds a temporal element to the changing of the geography of election manipulation in Russia. The choropleth maps generated for each of the 5 election years will allow us to compare the magnitude of round

[20] I explored turnout fraud and mapped it, but because the stochastic model that the 'spikes' package relies on is suitable only for vote share fraud estimation, turnout fraud has not been statistically tested in this project.

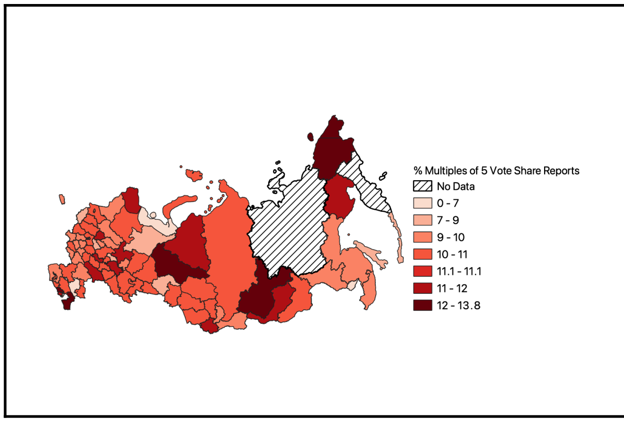
number manipulation committed in each individual region over the course of the past 20 years. This method reveals patterns in the spread and scale of election manipulation throughout Putin's regime, which may prove valuable in further understanding the relationship between elections and long-term authoritarian regimes.

## Results & Analysis

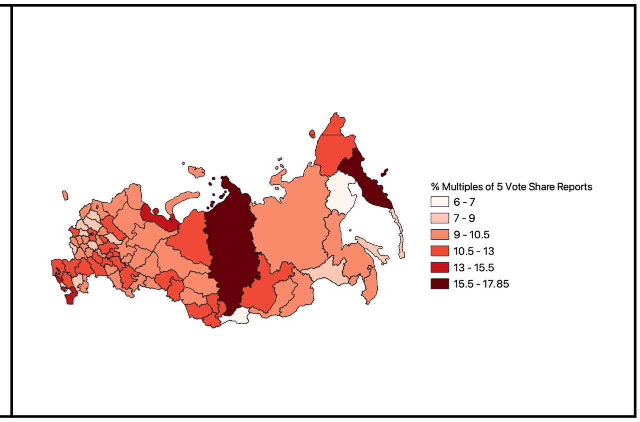
### *Data Visualization*

The maps below are rendered with natural (Jenks) breaks. Jenks breaks help visualize data clustering by minimizing each class' difference from the mean while differentiating each class from others. In other words, natural breaks maximize the variance of each class from other classes while minimizing the variance from the group mean. The legend values are in percent of total precincts in the region. For example, a value of 11.2 corresponds to 11.2% of the precincts in that administrative subject reporting multiples of 5 for vote share. Jenks breaks proved to be a much better method of visualizing the distribution of the density of reported multiples of 5 for vote share. It should be noted that because the scale is not standardized across election years, each map should be viewed as a stand alone that depicts where the highest density of round number reporting was located.

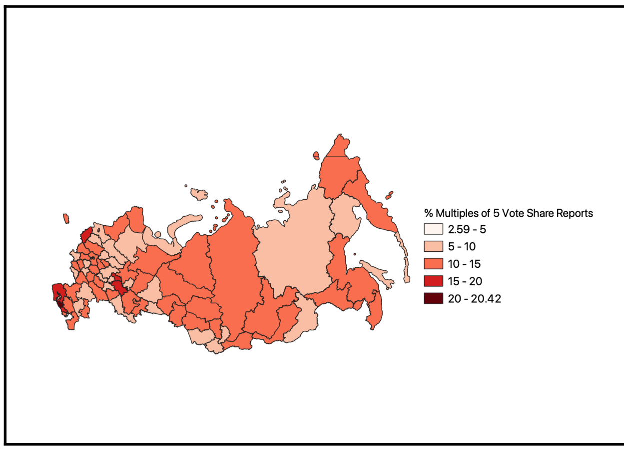
Notably, ethnic republics consistently have high percentages across all the available years of results. In particular, Dagestan is an influential outlier for both 2008 and 2012. In 2018 there is an uptick of eastern administrative regions that show higher volumes of round percentage numbers.



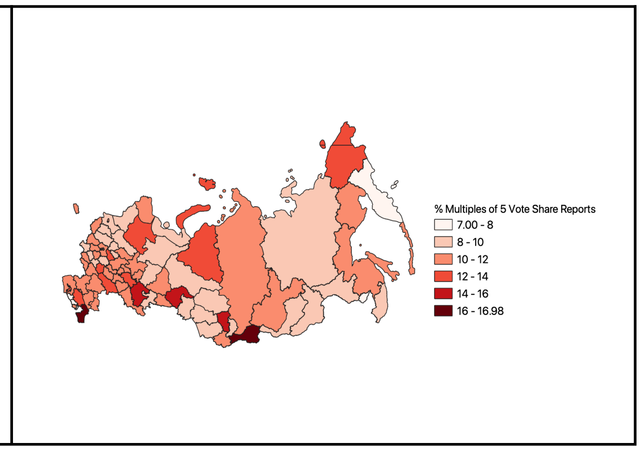

2000 [21]



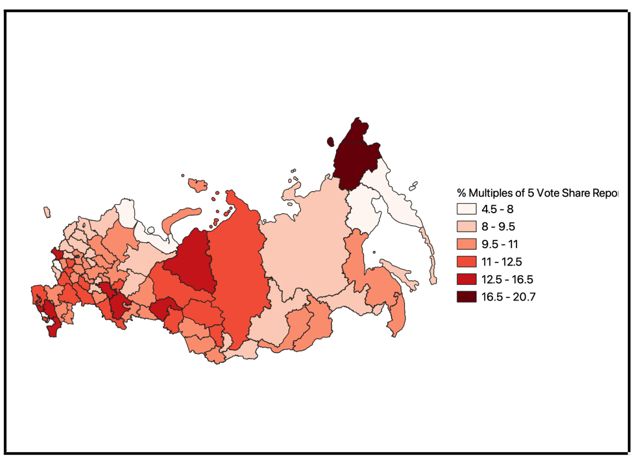

2004




2008




2012




2018

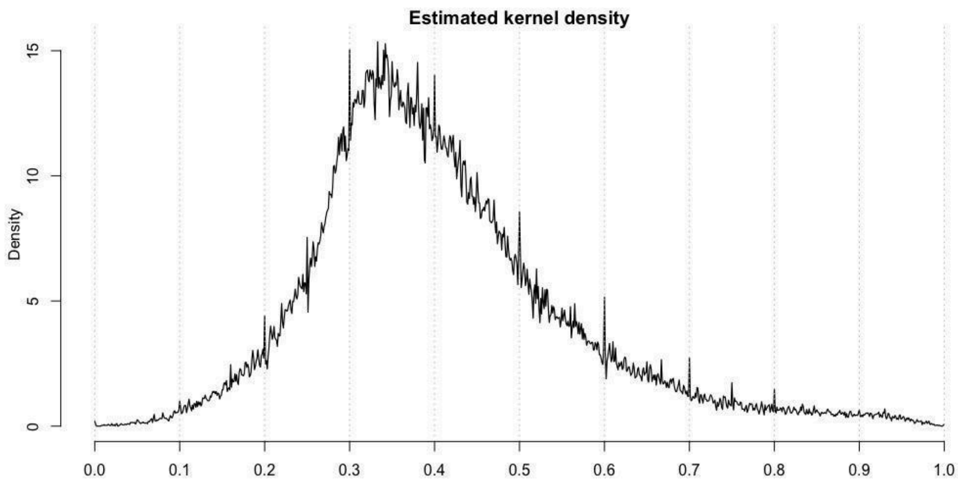
[21] No data for the Republic of Sakha and Chukotskiy Autonomous Okrug.

*Kernel Density Estimates*

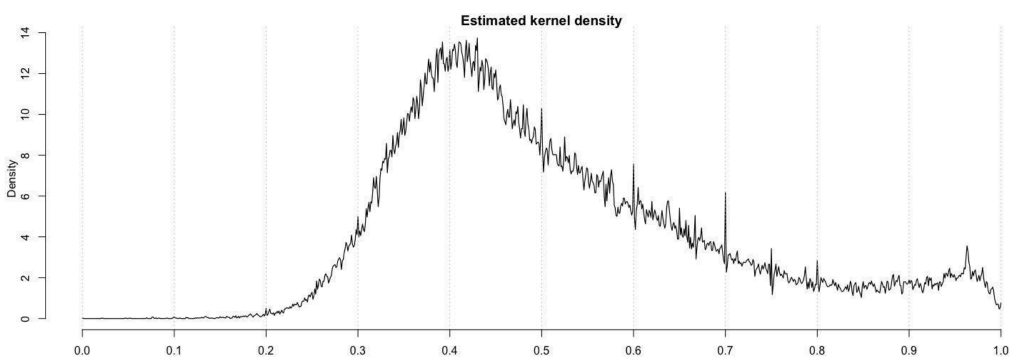
The Kernel Density Estimates for each year show the density probability distribution of vote share reported. The spikes show the density of precincts across Russia that reported that percentage in each election year. Not only does the shape of the distribution change and become left-skewed over the years, but the height of the spikes also grows from 2000 to 2018, indicating that there is a higher volume of precincts reporting those exact numbers and therefore a higher volume of precincts engaging in election result manipulation. There is also a visible increase in high percentage reporting precincts starting in 2004, resulting in a non-normal distribution, which is especially visible in 2008 as a bump in the distribution around 90-100%. The manipulation has also apparently skewed the overall distribution over the years. In 2000, the majority of precincts reported vote share between 30-40%. In subsequent years that shifts to 40+%. In 2018 the mode appears to be situated at 50%.

Considering the political climate throughout, the distribution changes make sense. Putin came to power in the year 2000, when fraud was at its relative minimum. Myagkov, Ordeshook, and Shakin speculate that this is due to the fact that regional bosses are uncertain about how long the newcomer will stay in office and are hesitant to be over eager with their loyalties and allegiances. There is an uptick in 2004 that signifies the confidence that some regional bosses have gained from 4 years of a Putin presidency. In 2007, Putin stepped down to make way for Medvedev's presidential campaign. Possibly as a result of concern surrounding Medvedev's performance, the candidate receives a boost reflected in the bump around 90% and obvious spikes at 60% and 70% vote share. A comparison of 2012 and 2018, demonstrates a dramatic shift into the higher ranges of vote share. The 2018 change could be a reaction to the uptick in opposition to the Kremin. Even with Navalny disqualified and removed from the race in 2016, it is possible that the Kremlin panicked that an oppositional candidate such as Navalny would even be able to gather a platform and supporters around him and as a result this was reflected in the vote share data from the 2018 elections.

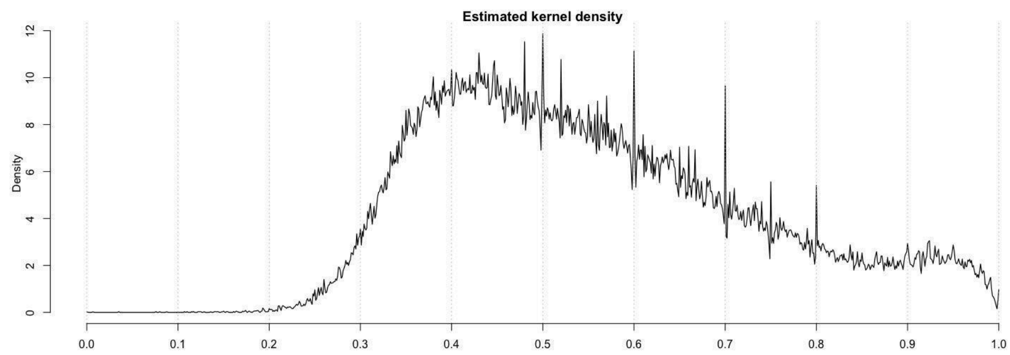
2000:



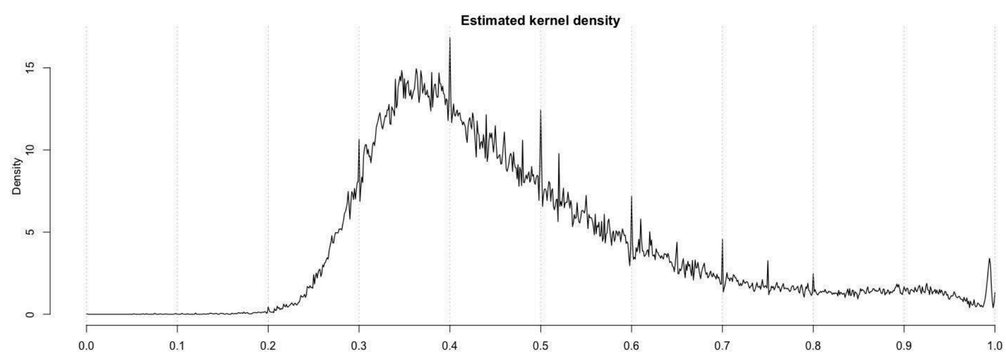

2004:



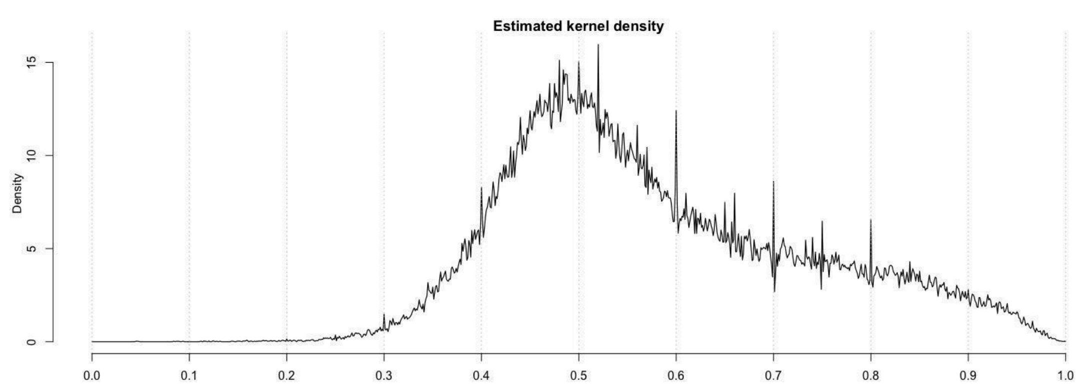

2008:


Estimated kernel density
Density

2012:


Estimated kernel density
Density

2018:


Estimated kernel density
Density

*Variance*

The results below show the statistical summaries of the estimated contamination (using the Rozenas' spikes package) for each election year. The summaries of the estimated fraud results reflect the changes that are visually apparent from the Kernel Density Estimates. From the year 2000 to 2018, the mean, standard deviation, variance, skewness, and kurtosis (spikiness) all increase. In other words, the distribution of the vote share becomes remarkably less normally distributed. Moreover, the increase in kurtosis, skew, and variance indicates what Myagkov, Ordeshook, and Shakin call the "bifurcation of the electorate."[22] That is, we are seeing the data become "noisier" because one subsection of the electorate is reporting increasingly fraudulent results. These places have had a rise in the amount of falsification of election results while most other places have remained at the same base level, leading to a greater difference between these two groups, which is reflected in the increase in the variance.

A normal kurtosis value for a normally distributed data set is around 3. A normal value for skew (indicating symmetrical data distribution) is 0. For all of the election years, skew is above normal and the distribution is leptokurtic. It is significantly above what we would expect for a normal value. The skew is also concerning. As is evident from the statistical summaries and the kernel density estimates, the entire distribution of vote share has shifted towards higher values from the year 2000 onwards. Medvedev's election year in particular saw a massive jump in fraudulent vote share reporting, which is reflected in the percentile distributions and skew.

[22] Myagkov, Mikhail, Peter Ordeshook, and Dimitri Shakin. *The Forensics of Election Fraud: Russia and Ukraine*.

EstFraud

| | Percentiles | Smallest | | |
|---|---|---|---|---|
| 1% | 0 | 0 | | |
| 5% | 0 | 0 | | |
| 10% | 0 | 0 | Obs | 82 |
| 25% | 0 | 0 | Sum of wgt. | 82 |
| 50% | 0 | | Mean | .3118293 |
| | | Largest | Std. dev. | .5550734 |
| 75% | .5 | 1.31 | | |
| 90% | .96 | 1.69 | Variance | .3081065 |
| 95% | 1.14 | 2.75 | Skewness | 2.753337 |
| 99% | 2.95 | 2.95 | Kurtosis | 11.95211 |

Summary of 2000 Vote Share Fraud Results

EstFraud

| | Percentiles | Smallest | | |
|---|---|---|---|---|
| 1% | 0 | 0 | | |
| 5% | 0 | 0 | | |
| 10% | 0 | 0 | Obs | 83 |
| 25% | 0 | 0 | Sum of wgt. | 83 |
| 50% | .15 | | Mean | .6540964 |
| | | Largest | Std. dev. | 1.463094 |
| 75% | .59 | 4.09 | | |
| 90% | 1.35 | 6.92 | Variance | 2.140644 |
| 95% | 2.6 | 7.46 | Skewness | 3.635268 |
| 99% | 7.49 | 7.49 | Kurtosis | 16.29205 |

Summary of 2004 Vote Share Fraud Results

EstFraud

| | Percentiles | Smallest | | |
|---|---|---|---|---|
| 1% | 0 | 0 | | |
| 5% | 0 | 0 | | |
| 10% | 0 | 0 | Obs | 83 |
| 25% | 0 | 0 | Sum of wgt. | 83 |
| 50% | .2 | | Mean | 1.586386 |
| | | Largest | Std. dev. | 5.024001 |
| 75% | .85 | 8.33 | | |
| 90% | 4.43 | 10.89 | Variance | 25.24058 |
| 95% | 7.68 | 11.71 | Skewness | 6.337533 |
| 99% | 41.37 | 41.37 | Kurtosis | 48.9895 |

Summary of 2008 Vote Share Fraud Results

EstFraud

| | Percentiles | Smallest | | |
|---|---|---|---|---|
| 1% | 0 | 0 | | |
| 5% | 0 | 0 | | |
| 10% | 0 | 0 | Obs | 83 |
| 25% | 0 | 0 | Sum of wgt. | 83 |
| 50% | .42 | | Mean | .9637349 |
| | | Largest | Std. dev. | 1.791709 |
| 75% | .95 | 4.49 | | |
| 90% | 2.65 | 5.85 | Variance | 3.210221 |
| 95% | 4.13 | 9.39 | Skewness | 3.363693 |
| 99% | 10.16 | 10.16 | Kurtosis | 15.72563 |

Summary of 2012 Vote Share Fraud Results

EstFraud

| | Percentiles | Smallest | | |
|---|---|---|---|---|
| 1% | 0 | 0 | | |
| 5% | 0 | 0 | | |
| 10% | 0 | 0 | Obs | 83 |
| 25% | 0 | 0 | Sum of wgt. | 83 |
| 50% | .18 | | Mean | 1.1 |
| | | Largest | Std. dev. | 3.015112 |
| 75% | .93 | 6.05 | | |
| 90% | 2.38 | 6.21 | Variance | 9.0909 |
| 95% | 5.91 | 9.12 | Skewness | 5.792106 |
| 99% | 24.03 | 24.03 | Kurtosis | 42.21342 |

Summary of 2018 Vote Share Fraud Results

**Conclusion**

Overall, the amount of estimated fraudulent election results reported in Russia increased throughout Putin's regime. From the resampled kernel density estimates and the statistical summaries of each election year, it is evident that the scale of fraud has increased across the board, but especially in some of the majority ethnic minority regions. Ethnic republics and autonomous regions consistently report high percentages of multiples of 5 for vote share (and turnout) results, but have upped the ante in the most recent election years as compared to the beginning of Vladimir Putin's regime. This lends credibility to Myagkov, Ordeshook, and Shakin's argument about the bifurcation of the Russian electorate, in which there is a growing gap between the amount of fraud committed by different regions.

The shifting of the vote share distribution towards higher values over the years demonstrates that there is such a high volume of election manipulation happening that the average has begun moving into a higher range. This is mostly due to the outliers that are the ethnic republics. For the most part, non-majority minority regions in Russia report consistent levels of election fraud throughout Putin's regime. However, it is these outliers that have shifted so dramatically towards higher percentage results that the entire distribution of the results has shifted as well. In the year 2000, Putin's mean vote share was approximately 35%. By 2018, it was just above 50%. In 2008, during Medvedev's election cycle, anxiety about his ability to perform at the polls might have created the extra "bump" at higher vote share percentages, causing the data to look almost like it is trending toward bimodality. This can be accounted for with some of the abnormally high fraud percentages reported by ethnic republics in that year. This is corroborated through an application of Rozenas' spikes package to the 2008 data. The

Republic of Ingushetia alone was estimated to have a 41% contamination rate in it's reported vote share.

This paper leaves room for exploration. While I originally set out to provide a visual representation of where the majority of fraud is happening, it does not test for spatial autocorrelation or estimate whether it is possible that regions "infect" their neighbors with inflated vote share results. This thesis has provided some clarity as to the general trends in Russian election manipulation, but it is by no means exhaustive.

**Glossary of Useful Terms**

*Churov's Saw:* A statistical phenomenon describing patterns in election data that point to results manipulation. Graphically, it appears as a series of saw-like or comb-like spikes that represent the volume or frequency of a specific reported result.

*Course integer rate:* Rate of whole number results reported for turnout or vote share.

*Federal Subject/Region:* One of 85 federal subjects of the Russian Federation. The largest administrative divisions in Russia. Equivalent to states in the U.S.

*Oblast:* A type of region/federal subject that is often ethnically homogenous.

*Pamfilova Peak:* A particular integer standout in election results. See Churov's Saw.

*Raion:* An administrative division of a federal subject.

*Republic/Ethnic Republic:* A type of region/federal subject that is often majority minority ethnic group.

*Round integer rate:* Rate of round (multiples of 5) integer results reported for turnout or vote share.

*Tik:* District ID. Abbreviation for Territorial Election Commission. Roughly 2,778 districts total in Russia.

*Types of federal subjects:* Oblasts, Krais, Republics (substantial ethnic minority), Autonomous Okrugs (substantial ethnic minority), Federal Cities (Moscow, Saint Petersburg, and, as of 2014, Sevastopol, which is disputed)

*Uik:* Precinct ID. Abbreviation for Local Election Commission. Roughly 95,000 total in Russia.

**Acknowledgments**

None of this would have been possible without the help of some amazing people that I am privileged to know. Thank you, first and foremost, to my second readors, mentors, and advisors along the way. Your guidance and willingness to answer all my questions and meet even for the smallest issues throughout this process has been a nurturing source of support. Thank you to the researchers who paved the way before me and painstakingly created the datasets and packages I have used. Without their foundational work, none of this would have been possible. Finally, thank you to my wonderful support base of friends who patiently listened to me re-explain my methodology to them over and over again and received endless text messages when things were not going so well.

## Appendix

### *Codebook*

region = region
tik = district ID
uik = precinct ID
v1 = total registered voters
v2 = total votes cast
v3 = total number of valid ballots
v4 = total number of invalid ballots
v5 = total number of discarded ballots
Putin = number of total votes cast for Putin
n1 = number of districts (tik) in each region
n2 = frequency of unique tik
turnout = ((v3 + v4)/v1)*100
turnout_rounded = turnout rounded to 0.5
m1 = residuals of turnout_rounded/5
m2 = m1 where value==1 if turnout_rounded is a multiple of 5, ==0 if not
m3 = proportion of round integer precincts by region (mean m2)
m4 = proportion of round integer precincts by district (tik) (mean m2)
voteshare = percent vote share for Putin (Putin/(v3 +v4))*100
voteshare_rounded = vote share rounded to 0.5
k1 = voteshare_rounded/5
k2 = k1 where value==1 if voteshare_rounded is a multiple of 5, ==0 if not
k3 = proportion of round integer precincts by region (mean k2)
k4 = proportion of round integer precincts by district (tik) (mean k2)
ratio = abs(m4/k4)

### *Stata Replication Code*

```
generate voteshare = (Putin/(v3 +v4))*100
generate turnout = ((v3 + v4)/v1)*100
keep if v2>50
keep if turnout<100
drop if inlist(region, "99 Территория за пределами РФ")
drop if inlist(region, "город Севастополь")
sort region
by region: generate n1 = _N
sort tik
by tik: generate n2 = _N
gen turnout_rounded=round(turnout,.5)
generate m1 = mod(turnout_rounded, 5)
generate m2= 1 if m1==0
replace m2=0 if m1>0
sort region
by region: sum m2
```

```
bysort region: egen m3 = mean(m2)
sort tik
bysort tik: egen m4 = mean(m2)
gen voteshare_rounded=round(voteshare,.5)
generate k1 = mod(voteshare_rounded, 5)
generate k2= 1 if k1==0
replace k2=0 if k1>0
sort region
by region: sum k2
bysort region: egen k3 = mean(k2)
sort tik
bysort tik: egen k4 = mean(k2)
generate ratio = abs(m4/k4)
order v1 v2 v3 v4 v5 Putin n1 n2 turnout turnout_rounded m1 m2 m3 m4 voteshare
voteshare_rounded k1 k2 k3 k4 ratio, after (uik)
sort region tik uik
egen id = group(region)
```

***R Replication Code***

```
### set library ###
library(readxl)
library(dplyr)
library(VGAM)
library(emdbook)
library(BenfordTests)
library(spikes)
library(haven)
library(parallel)
library(foreach)
library(doParallel)

setwd("~/Desktop/BA Thesis/crash tests/R Data")

--------------------------------
  ### 2000 ###

  setwd("~/Desktop/BA Thesis/crash tests/R Data/2000xlsx")

## import all 2000 regions ##
list.2000 <- list.files(path = "~/Desktop/BA Thesis/crash tests/R Data/2000xlsx")
regions2000 <- lapply(list.2000, read_excel)

## set column names ##
colnames <-  c("region", "N", "t", "v")
for (i in seq_along(regions2000)){
  colnames(regions2000[[i]]) <- colnames
```

```
}

## generate 2000 results ##
# results_2000 <- lapply(regions2000, spikes)
tryCatch({
  results_2000 <- lapply(regions2000[1:82], spikes)
}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})


--------------------------------
  ### 2004 ###

  setwd("~/Desktop/BA Thesis/crash tests/R Data/2004xlsx")

## import all 2004 regions ##
list.2004 <- list.files(path = "~/Desktop/BA Thesis/crash tests/R Data/2004xlsx")
regions2004 <- lapply(list.2004, read_excel)

## set column names ##
for (i in seq_along(regions2004)){
  colnames(regions2004[[i]]) <- colnames
}

## generate 2004 results ##
# results_2004 <- lapply(regions2004, spikes)
tryCatch({
  results_2004 <- lapply(regions2004[1:83], spikes)
}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})

--------------------------------
  ### 2008 ###

  setwd("~/Desktop/BA Thesis/crash tests/R Data/2008xlsx")

## import all 2008 regions ##
list.2008 <- list.files(path = "~/Desktop/BA Thesis/crash tests/R Data/2008xlsx")
regions2008 <- lapply(list.2008, read_excel)

## set column names ##
for (i in seq_along(regions2008)){
  colnames(regions2008[[i]]) <- colnames
}

## generate 2008 results ##
# results_2008 <- lapply(regions2008, spikes)
tryCatch({
```

```
  results_2008 <- lapply(regions2008[1:83], spikes)
}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})


--------------------------------
  ### 2012 ###

  setwd("~/Desktop/BA Thesis/crash tests/R Data/2012xlsx")

## import all 2012 regions ##
list.2012 <- list.files(path = "~/Desktop/BA Thesis/crash tests/R Data/2012xlsx")
regions2012 <- lapply(list.2012, read_excel)

## set column names ##
for (i in seq_along(regions2012)){
  colnames(regions2012[[i]]) <- colnames
}

## generate 2012 results ##
# results_2012 <- lapply(regions2012, spikes)
tryCatch({
  results_2012 <- lapply(regions2012[1:83], spikes)
}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})

--------------------------------
  ### 2018 ###

  setwd("~/Desktop/BA Thesis/crash tests/R Data/2018xlsx")

## import all 2018 regions ##
list.2018 <- list.files(path = "~/Desktop/BA Thesis/crash tests/R Data/2018xlsx")
regions2018 <- lapply(list.2018, read_excel)

## set column names ##
for (i in seq_along(regions2018)){
  colnames(regions2018[[i]]) <- colnames
}

## generate 2018 results ##
# results_2018 <- lapply(regions2018, spikes)
tryCatch({
  results_2018 <- lapply(regions2018[1:83], spikes)
}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
```

***QGIS Projection Info***

EPSG:3576

SR-ORG:8568