CEIS TOR VERGATA Research Papers

Università di Roma Tor Vergata

ISSN 2610-931X

CEIS Tor Vergata

RESEARCH PAPER SERIES

Vol. 22, Issue 5, No. 584 – October 2024

With a Little Help From the Crowd: Estimating Election Fraud with Forensic Methods

Christoph Koenig

With a Little Help From the Crowd: Estimating Election Fraud with Forensic Methods

Christoph Koenig *

October 28, 2024

Abstract

Election forensics are a widespread tool for diagnosing electoral manipulation out of statistical anomalies in publicly available election micro-data. Yet, in spite of their popularity, they are only rarely used to measure and compare variation in election fraud at the sub-national level. The typical challenges faced by researchers are the wide range of forensic indicators to choose from, the potential variation in manipulation methods across time and space and the difficulty in creating a measure of fraud intensity that is comparable across geographic units and elections. This paper outlines a procedure to overcome these issues by making use of directly observed instances of fraud and machine learning methods. I demonstrate the performance of this procedure for the case of post-2000 Russia and discuss advantages and pitfalls. The resulting estimates of fraud intensity are closely in line with quantitative and qualitative secondary data at the cross-sectional and time-series level.

[Word count abstract: 148]

[Word count manuscript: 9,434]

Keywords: Bayesian Additive Regression Trees, Election Forensics, Election Fraud, Election Monitoring, Machine Learning, Russia

*Assistant Professor, University of Rome Tor Vergata, Department of Economics and Finance, Via Columbia 2, 00133 Roma, Italy and Honorary Research Fellow, University of Bristol, School of Economics, The Priory Road Complex, Priory Road, BS8 1TU Clifton, United Kingdom. Email: Christoph.Koenig@uniroma2.it. I would like to thank Michael Bechtel, John Marshall, Matthew Pratola, Alfonso Russo, David Szakonyi and Andreas Wiedemann for important discussions and useful insights, Dmitry Kobak and Sergey Shpilkin for help with the Russian election micro-data, Maxim Krukov for sharing the GOLOS reports and Adam Kapelner and Walter Mebane for guidance on the R packages `bartMachine` and `eforensics`, respectively. All remaining errors are mine.

1 Introduction

The last decades have seen a continued rise in quantitative analyses of election fraud.[1] While many studies have sought to explain differences between countries, recent research is increasingly focusing on the drivers of manipulation at the sub-national level within a single country.[2] A significant challenge of such analyses is the lack of reliable data. The perpetrators of election fraud, understood here as "clandestine and illegal efforts to shape election results" (Lehoucq, 2003), naturally try hiding their actions. Furthermore, the data sources used in country-level datasets such as newspaper articles and election observer reports typically do not have sufficient coverage across sub-national units. Forensic methods, in theory, provide a way to circumvent these issues as they allow generating evidence for or against fraud in the above sense out of statistical anomalies in publicly available election micro-data with often universal coverage. The potential of these methods, however, is typically limited by 1) the wide array of available indicators a researcher may choose from, 2) the potential variation in manipulation methods applied across localities and elections and 3) that forensic indicators as such typically do not constitute a comparable measure of fraud intensity.

This article proposes a context-independent, data-driven procedure seeking to overcome the above problems. The key novelty is leveraging data from crowd-sourced election monitoring (CSEM) to train a machine learning (ML) model which flexibly approximates the relationship between observed instances of fraud and a set of forensic indicators for

[1] For an overview, see the articles by Lehoucq (2003), Gandhi and Lust-Okar (2009) and Mares and Young (2016) as well as the books by Alvarez et al. (2008) and Simpser (2013).

[2] See, for instance, Callen and Long (2015), Sjoberg (2014) and Cantú (2019).

the country of interest.[3] One can then use the trained model to diagnose irregularities and estimate fraud also for other parts of the country or time periods where only election micro-data is available. I demonstrate the procedure with data from post-2000 Russia and discuss advantages and pitfalls. Since election micro-data is getting increasingly easier to access (Rueda et al., 2023) and CSEM is becoming more widespread through increased internet access in developing countries (Grömping, 2017), my approach is potentially applicable to several other contexts with limited data.

Methodologically, I follow earlier work at the country-level by Montgomery et al. (2015a) which makes use of *Bayesian Additive Regression Trees* (BART) – an ensemble ML method combining elements of decision trees and boosting. For my analysis of Russia, I obtained about 5,600 reports of observed electoral law violations during the elections 2011 and 2012 from GOLOS, an independent election monitoring NGO. Using BART allows me to estimate the relation between reported fraud and 14 forensic indicators from the relevant literature with a high degree of flexibility and without imposing any functional form. Having trained the model, I can predict fraud for each of Russia's roughly 2,700 districts in all 10 national elections between 2000 and 2021 using only forensic indicators calculated from official election micro-data.

Accurately quantifying the extent of election fraud across an entire country over several elections based on official election data and a few thousand fraud reports for two elections is necessarily associated with some challenges. The most important ones are related to limitations of the fraud report data. For instance, reporting was very high in metropolitan areas, whereas the vast majority of the country, especially the notoriously fraud-ridden rural parts in the Southwest, often did not even record a single act of fraud. As a result,

[3] CSEM are digital platforms set up by activists or non-governmental organizations (NGOs) which allow citizens to collect and share evidence of manipulation.

the ML algorithm may wrongly end up predicting knowledge or usage of GOLOS instead of electoral manipulation. Furthermore, a very skewed distribution of outcomes in the training data, as in this case, may lead to over-predicting the majority class.

I address the above issues in several ways: first, I only use information on whether districts detected any fraud at all instead of report frequency. Second, I establish outcome balance in the training data by applying the Synthetic Minority Oversampling Technique (SMOTE) by Chawla et al. (2002). Finally, and most importantly, I create three alternative sets of predictions by restricting the training data in different ways with the aim of excluding mis-classified observations: the first one exploits that many reports did not concern fraud in the sense of hidden, illegal actions aimed at directly influencing the results but more innocuous violations of electoral law such as the presence of campaign materials in polling stations. Comparing only *Reporting* districts with an actual fraud report to those with only a non-fraud report should purge many false negatives from the data. The second restricted sample includes only *Switcher* districts which reported fraud in either the 2011 or the 2012 election but not the other one. This isolates variation *within* areas and assures that fraud status is independent of fixed area characteristics. The last restricted sample combines the two previous approaches and only makes use of *Reporting Switchers*.

While the internal validity of the different BART models can be assessed by comparing predictions and test data, this necessarily assumes the absence of mis-classified observations. Benchmarking the models' actual performance thus is not feasible without making use of reliable, secondary data. The most useful information in this regard are scores of electoral fairness from Petrov and Titkov (2013) for a cross-section of 83 regions, the next-highest unit in Russia's administrative system, and a time-series of country-level

fraud estimates based on the NELDA dataset by Hyde and Marinov (2012). Since none of these series is available at the district-level, I first aggregate predictions and then benchmark them against the aforementioned secondary sources by comparing their bivariate correlations. I find that, overall, the *Reporting Switcher* sample outperforms the other strategies. Using this sample, I also provide fraud estimates at the regional level from 2000 to 2021 and document a high degree of congruence between with qualitative accounts of electoral manipulation from the Organization for Security and Co-operation in Europe (OSCE) and other sources. Omitting the country's two largest cities, Moscow and St. Petersburg, results in very similar estimates.

Finally, in the last part of the analysis, I seek to understand which forensic indicators have the highest predictive power and whether their relationship with reported fraud is in line with their theoretical premises. A straightforward metric to address the first question are inclusion rates, i.e. the share of a full BART model's splitting rules in which a particular variable is used. I find that the most relevant forensic indicators in my setup are the *resampled kernel density* method (RKD) by Rozenas (2017), the *integer percentage* approach (IP-TURN) by Kobak et al. (2016a) and the *excess turnout-vote share correlation* (TVSC-XS) popularized by Myagkov and Sobyanin (1996). For the second question, I look at partial dependence (PD) plots which show the average outcome prediction at specific quantiles of the three aforementioned variables. The graphs reveal that, while all three variables are roughly in line with theoretical expectations, the effect of IP-TURN also exhibits some contradictory patterns.

This paper makes two main contributions. The methodological one is presenting a procedure which leverages ML methods, CSEM data and election forensics to estimate electoral manipulation for a particular country at the sub-national level using only election

micro-data. This procedure is an extension of Montgomery et al. (2015a), which is also the most closely related work in this area. Their analysis uses BART to estimate the relationship between a national-level fraud measure constructed from the NELDA dataset and several forensic indicators as well as contextual risk factors. My study also draws on pioneering work by Cantú and Saiegh (2011) who train a Naive Bayes classifier on synthetic data containing instances of vote stealing to detect cases thereof in Buenos Aires between 1931 and 1941 using Benford's law (BL). Levin et al. (2016) and Zhang et al. (2019) apply Random Forests, another ML ensemble method, to synthetic data with instances of vote stealing and ballot stuffing. Having trained the model using turnout and vote shares as input, they predict the likelihood of local-level fraud in the 2013 and 2015 elections in Argentina.

My approach differs from this work in few aspects: first, I use crowd-sourced fraud reports as an outcome variable. This is because data sources like NELDA are not available for sub-national units in this and most other contexts.[4] I also do not use synthetic data but, rather, rely on forensic indicators, which implicitly compare realized election outcomes with a simulated benchmark. Second, I use a wide array of forensic tools as explanatory variables, covering 14 indicators in total and spanning turnout-inflating, number-manipulating as well as rounding fraud. Third, I do not use contextual variables to fit the BART model. This is less problematic in my case since I am analysing localities within a single country with a more uniform political context instead of comparing countries across the globe. In fact, many contextual variables could be related to an area's propensity to report fraud and may thus additionally bias predictions in that direction.

[4] There are few cases where primary, disaggregated evidence is available like as in the studies by Cantú (2019) and Callen and Long (2015). However, this evidence typically only concerns one particular type of fraud.

Furthermore, contextual variables are, in practice, often not available at the sub-national level.

The second, substantive contribution of this paper is providing comparable estimates of election fraud in Russia at the national and sub-national level for all country-wide elections from 2000 to 2021. Existing research has documented turnout-inflating (Myagkov et al., 2009; Enikolopov et al., 2013), number-manipulating (Skovoroda and Lankina, 2016) as well as rounding fraud (Kobak et al., 2016b; Rozenas, 2017). However, no work so far has sought to combine several forensic indicators into a single metric and measure election fraud across manipulation methods at the district- or regional level.[5] An advantage of my estimates is that they are designed to mimick voting patterns in districts where citizens have in fact reported fraud. Furthermore, they correlate strongly with near-ideal, aggregate secondary data and qualitative evidence.

My estimates naturally also have limitations. First, they assume that reporting and forensic indicators in the training sample are only linked through fraud and not a third variable that could bias the estimates. Second, they require that all fraud methods applied between 2000 and 2021 were also used in the training elections 2011/2012 and, at least to some extent, detected. Lastly, since the district-level predictions can only approximate whether fraud occurred, their regional aggregates can consequentially only measure the maximum share of votes potentially affected by manipulation. Hence, while my predic-

[5] To the best of my knowledge, there exists also no work for other countries. Leemann and Bochsler (2014) apply several number-based tests for a referendum in Switzerland but do not combine their results into single metric.

tions can give a good sense about the intensity of fraud, they cannot quantify how much election outcomes were actually shifted by it.[6]

This article proceeds as follows: Section 2 introduces the data sources, followed by Section 3 which provides an overview of the forensic indicators used in the analysis. Section 4 discusses the methodology. Section 5 presents the main empirical results and Section 6 evaluates the performance of the individual fraud indicators. Section 7 concludes.

2 Data

2.1 Election data

In this paper, I exclusively focus on *national* elections for the Russian president and the State *Duma* parliament. Their organization strongly corresponds to Russia's administrative divisions: the Central Election Commission (CEC) is the highest authority and coordinates Russia's 89 Regional Election Commissions (RECs).[7] The RECs, on the other hand, coordinate the work of the roughly 2,700 Territorial Election Commissions (TECs). The TECs' territories typically coincide with those of districts (rayons), which are the second-level of Russia's administrative divisions.[8] Finally, the organization of the individual voting stations is carried out by about 95,000 precinct election commissions (PECs). For my analysis I use publicly available PEC-level data for five presidential and five Duma elections in Russia from 2000 to 2021 compiled by Sergey Shpilkin from the

[6] Importantly, this caveat also applies to forensic indicators producing an estimate of fraud levels like the RKD measure by Rozenas (2017) and the Finite mixture model/eforensics approach in Mebane Jr. et al. (2022).

[7] I assume that Russia kept its original 2014 borders over the entire study period 2000 to 2021.

[8] In the case of larger cities and detached settlements, districts may host several TECs. For simplicity, I use the names district and TEC interchangeably.

Central Election Commission website.[9] The main variables of interest are the size of the electorate, turnout and votes for incumbent candidates and parties. The incumbent is defined as United Russia in parliamentary elections (2003, 2007, 2011, 2016 and 2021) and Vladimir Putin (2000, 2004, 2012 and 2018) and Dmitry Medvedev (2008) in presidential races.[10]

2.2 Fraud reports

The invitation of foreign election observers has become almost a standard practice since the 1990s, even for non-democratic regimes (Kelley, 2012). Domestic monitoring through a country's own citizens, on the other hand, is a more recent, yet fast-growing, phenomenon (Grömping, 2017). The global spread of internet access and smartphones has made the mobilization and coordination of observers substantially easier for domestic actors engaged in election monitoring. One particularly powerful innovation has been the usage of CSEM which allows virtually any citizen to become an election observer and submit evidence of manipulation via social media, text message, email or a website to activist groups and non-governmental organizations (NGOs). To give a broad idea about the scope and spread of CSEM, Table 1 provides a non-exhaustive overview of its application in recent elections. While a comprehensive evaluation is still missing, anecdotal evidence suggests that CSEM may indeed be an efficient tool at deterring, or at least displacing, fraudulent activities

[9] The data is nearly complete for all elections, apart from 2000 where no PEC data is available for the Republic of Sakha and Chechnya. For further details, see Shpilkin (2021) and Kobak (2023).

[10] Turnout is defined as the sum of votes cast over electorate size. For calculating the indicator by Kobak et al. (2016b), I apply the turnout definition used in their paper. I also follow their example by considering only the proportional votes in 2003, 2016 and 2021 when State Duma members were elected in a mixed system with half of the 450 seats allocated via majoritarian single-member districts and the other half proportionally through regional party lists.

Table 1: Examples of CSEM around the world

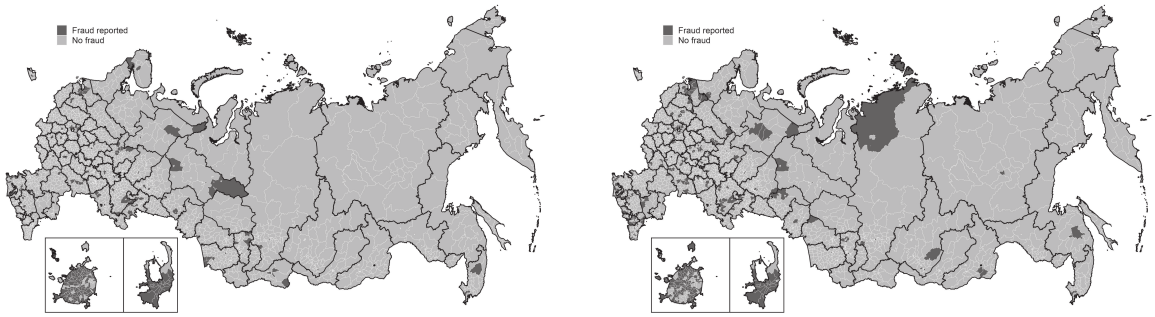
Country	Year	Indicative references	Name of crowd-sourcing platform
Armenia	2013	Vardanyan (2013)	Iditord
Guinea	2010	Bott et al. (2014)	Guinée Vote 2010 Témoin
Honduras	2013	Arias et al. (2015)	VotoSocial
Indonesia	2019	Gunawan and Ruldeviyani (2020)	KawalPilpres
Kenya	2008	Ajao (2022)	Ushahidi
Kenya	2013	Ajao (2022)	Ushahidi
Mexico	2009	Salazar and Soto (2011)	¡Cuidemos El Voto!
Nigeria	2011	Bailard and Livingston (2014)	ReVoDa
Russia	**2011**	**Bader (2013)**	**GOLOS**
Russia	**2012**	**Bader (2013)**	**GOLOS**
Uganda	2011	Hellström (2015)	UgandaWatch
Ukraine	2012	Herron and Sjoberg (2016)	Maidan-Monitoring
Ukraine	2014	Herron and Sjoberg (2016)	Maidan-Monitoring
Tanzania	2015	Shayo (2021)	Uchaguzi Wetu

(Grömping, 2017). My empirical results demonstrate another, so far overlooked, benefit of CSEM: helping to identify fraud in non-monitored areas and elections.

In my analysis, I use fraud reports collected by GOLOS, an independent Russian NGO specialized in election monitoring.[11] During the 2011 parliamentary and the 2012 presidential elections, the association ran the CSEM project *Karta Narusheniy* (map of violations) which provided a platform for citizens to anonymously report incidents of fraud and send reports of observed electoral law violations via phone, internet, and text message. Users could also provide information about the time and type of irregularity observed.[12] Of particular interest for this analysis are categories of election-day irregularities closely linked to fraud in the sense of "clandestine and illegal efforts to shape election results" (Lehoucq, 2003): 1) violation of observers' rights (incl. those of committee members and media), 2) illegal voting (incl. irregularities in home or absentee votes) and 3) counting irregularities (or other aspects of falsely processing of results). Lastly, the reporters could also provide the location where the action was witnessed which allows matching them to districts.

[11] My study is not the first to use this data source. See, for instance, Bader (2013), Bader and Schmeets (2013) and Skovoroda and Lankina (2016)

[12] For a more detailed description of the *Karta Narusheniy* data, see Bader (2013).



A. Reported fraud 2011 B. Reported fraud 2012

FIGURE 1

Notes: Maps of the Russian Federation showing reported fraud in the elections 2011 and 2012 across regions (thick lines) and districts (thin lines). White areas with thick black borders denote missing data.

2.3 Caveats

While crowd-sourced fraud reports are a highly valuable data source, they also have important drawbacks. The key issue is under-reporting which may stem from social pressure, fear of retaliation, lack of information or technical issues. Fraud may have also been more difficult to detect in rural areas than in densely populated cities. Consistent with this, 43% of election-day irregularities in 2011 and 26% in 2012 came from Moscow and St. Petersburg rather than the the notoriously fraud-ridden ethnic republics in the Southwest (Lukinova et al., 2011). Consequently, I do not use the frequency of reports but a binary variable *Fraud* which classifies a specific district-election cell as manipulated if any fraudulent action has been reported. This new binary variable thus captures the extensive rather than the intensive margin of fraud, i.e. the difference between final election outcomes and voters' actual choices on election day. Furthermore, Moscow and St. Petersburg still make up about 25% of the district-election cells reporting fraud. I will revisit this issue in Section 5.1.

Comparing only TECs with and without a report, however, does not suffice to compensate the data's shortcomings as illustrated by Figure 1 which shows the geographic

distribution of *Fraud* across districts in 2011 and 2012. Even though both elections were widely denounced for their high degree of manipulation, no single fraud report was filed in the vast majority of districts (84% in 2011, 81% in 2012). Those districts with a fraud report, were predominantly located in urban areas, in particular the cities of Moscow and St. Petersburg depicted in the bottom left of each map (OSCE, 2012b,a). Fraud reports were also more likely to come from the more densely populated Western part of the country rather than the more rural districts in the Far East.[13]

These patterns point towards two general issues in the prediction of election fraud highlighted by Cantú and Saiegh (2011): first, the number of *clean* cases typically strongly outnumbers the manipulated ones. Training an ML model on such data is likely to lead to predictions biased towards correctly predicting the majority class. The second is the high degree of mis-classified cases which may induce an ML model trained to predict fraud to, mistakenly, predicting correlates of reporting, such as development and urbanization, or of particular forensic measures.[14] I tackle these concerns in two ways: first, I use SMOTE to balance the distribution of outcomes in the training data. Second, I train the BART model also on three restricted samples which, to varying degrees, attempt to purge potentially mis-classified observations from the training data.

[13] In fact, Skovoroda and Lankina (2016) show that reports are correlated with a regions' education level and distance from Moscow.

[14] Rozenas (2017) shows that the occurrence of natural spikes in the vote share distribution at focal percentages decreases in the amount of voting stations. The TVSC indicator, on the other hand, relies on area homogeneity and could thus also be influenced by voters' geographical segregation.

2.4 Secondary data

To assess how well my fraud estimates capture cross-sectional variation across regions, I use data by Petrov and Titkov (2013) on regions' electoral fairness between 2006—2010 based on experts' assessments. I use the inverse of the original variable to impose a positive correlation with reliable proxies of fraud intensity. I also create a national-level time-series of manipulation levels in Russia based on the National Elections Across Democracy and Autocracy (NELDA) dataset by Hyde and Marinov (2012) which provides detailed information on the context, proceeding and outcomes of all national elections across the world. To construct a measure of electoral manipulation, I follow Montgomery et al. (2015a) and aggregate seven variables particularly related to fraud into a single measure through a standard three-parameter item response theoretic (IRT) model estimated on the full dataset. Using the most recent NELDA version 6.0 from 2021 provides information for Russia up to the 2018 election (Hyde and Marinov, 2021).

3 Forensic indicators

3.1 Selection criteria

For selecting the forensic indicators evaluated in the BART model, I applied the following criteria. First, the indicator should be *forensic*, i.e. seek to detect a statistical anomaly, rather than merely an extreme election outcome like very high turnout or incumbent vote share. Second, for very similar forensic tools, I selected the more recent or empirically tested one. Third, when different variants of an indicator were offered by the authors and none of them was ex-ante clearly superior or inferior, I typically included all of them.

3.2 Simulations

Several indicators in this section are using a Pearson's chi-squared test to compare the discrete distribution of digits in election outcomes within a particular area to a theoretical benchmark such as BL. This approach is problematic in my setup: first, the theoretical distributions only hold asymptotically whereas the number of precincts within a particular TEC ranges between 1 and 438 with a median of 30. In addition, the assumptions of the Pearson's chi-squared test break down if the expectation for any of the ten digits is less than one or below five for more than two digits (Cochran, 1954). For the BL2-test described below, this would apply to all TECs with less than 57 precincts, which is almost twice the median.

To sidestep the above issues, I perform the TEC-level digit distribution tests as follows: first, I do not rely on theoretical counter-factual distributions but on Monte-Carlo simulations following Kobak et al. (2016b). More precisely, I draw for each precinct 10,000 simulated totals of incumbent votes or absolute turnout from a binomial distribution with the number of draws equal to absolute turnout or the electorate size and a success probability equal to the actual incumbent share or turnout rate. I then extract from all draws the relevant digits and derive simulated distributions over the digits 0 to 9 for each TEC in every election. Second, for TECs which do not satisfy the criterion defined by Cochran (1954), I use an exact multinomial test. In both cases, low p-values indicate a significant deviation between the two and a high likelihood of manipulation.[15]

[15] To speed up the computation of multinomial tests, I make use a recent algorithm by Resin (2023).

3.3 Turnout-vote share correlation (TVSC)

The TVSC is one of the oldest and most widely used tools for detecting election fraud in Russia (Myagkov and Sobyanin, 1996). The core idea is that, absent manipulation, the correlation between how many people vote and their choice across sub-units of a particular area should be close to the area's average support of the respective candidate or party. An implausibly high TVSC or one in excess of the candidate's vote share (TVSC-XS) are thus potential signs of turnout-inflating fraud like ballot-stuffing or multiple voting (Myagkov et al., 2009). Enikolopov et al. (2013) have also shown that the TVSC in Moscow voting stations 2011 decreased when election observers were randomly deployed. The TVSC's main prerequisite is area homogeneity and the absence of any other systematic correlation between turnout and candidate preferences. In my analysis, I use highly disaggregated data and exploit variation across PECs within individual districts, which is the most granular analysis possible with existing data. — *Fraud indicators used:* TVSC and TVSC-XS.

3.4 BL2 test

A widely used forensic tool is the BL2 (or 2BL) test which posits that, absent fraud, the second digits of election outcomes like incumbent votes and turnout numbers should follow Benford's distribution (Mebane Jr., 2008). However, Deckert et al. (2011) and Rozenas (2017) show that the BL2 test tends to produce false positives and Mebane Jr. (2015) demonstrates that the test may also be sensitive to strategic voting. Past work has applied the BL2 test to Russian elections in the 2000s but only at the national rather than the sub-national level (Mebane Jr., 2013). Following Section 3.2, the BL2-test in my analysis uses simulated distributions of second digits rather than the theoretical ones implied by

BL. — *Fraud indicators used:* p-value of BL2-test for incumbent vote (BL2-INC) and turnout (BL2-TURN).

3.5 Last-digit (LD) test

In a seminal article, Beber and Scacco (2012) argue that, under fairly general assumptions, the last digits in reported election outcomes should follow a uniform distribution. Due to human preferences for specific numbers, man-made or manipulated election outcomes should thus be tilted towards particular digits and significantly deviate from this benchmark. Skovoroda and Lankina (2016) applied a variant of this test to recent Russian elections with a focus on the excessive occurrence of zeroes in turnout counts across voting stations. For my analysis, I use simulated LD distributions as a benchmark and allow biases to vary across numbers and TECs.[16] — *Fraud indicators used:* p-value of LD test for incumbent vote (LD-INC) and turnout (LD-TURN).

3.6 Digit-distance (DD) test

Beber and Scacco (2012) also propose a forensic indicator based on the distance between the last and second-last digit. This test is motivated by proven biases in human random number generation which often result in too few digit repetitions and excessive use of adjacent digits. In theory, the absolute distance in the last two digits of an election outcome should follow that of two uniformly distributed numbers. To the best of my knowledge, the DD test has so far not been applied to Russian elections. Again, I use simulated DD distributions as a benchmark. — *Fraud indicators used:* p-value of DD test for incumbent vote (DD-INC) and turnout (DD-TURN).

[16] For an extension and experimental evaluation of the last-digit test, see Medzihorsky (2015) and Mack and Stoetzer (2019).

3.7 Integer percentages (IP)

Kobak et al. (2016b,a) look at human biases in the creation of relative numbers, i.e. percentages. More precisely, they document a rising amount of precincts in Russia's national elections 2000 to 2020 reporting turnout or incumbent vote shares with an integer percentage (IP) (Kobak et al., 2018, 2020). Following their definition, an IP is any reported percentage at most 0.05 percentage points away from an integer.[17] The underlying assumption is that manipulating election officials have a target percentage in mind and then tweak absolute numbers to match this as closely as possible. Their method compares the amount of actual IP precincts with those arising from Monte Carlo simulations. For my analysis, I use a one-sided t-test for the null hypothesis whether the simulated number of IP precincts could be larger than the observed ones. — *Fraud indicators used:* p-value of IP-test for incumbent vote (IP-INC) and turnout (IP-TURN).

3.8 Resampled kernel density (RKD)

A salient feature of Russia's recent election results are abnormal spikes at focal percentages like 60% or 75% in the distribution of turnout and incumbent party vote share across precincts. The RKD method proposed by Rozenas (2017) aims to quantify the extent of this phenomenon. To avoid the detection of statistical artefacts unrelated to fraud, the RKD method first calculates a smooth counter-factual distribution from the data before comparing it to the actual outcome. The share of precincts positively deviating from the counter-factual then provides an approximation of the extent of fraud. — *Fraud indicator used:* Estimated share of fraudulent precincts (RKD).

[17] Similar indicators of fraud have been used by Rundlett and Svolik (2016) and Kalinin (2022).

3.9 Finite mixture model/eforensics (FMM)

Similar to RKD, the finite mixture model proposed by Mebane Jr. et al. (2022) tries to infer the share of fraudulent precincts within a given locality from distributional anomalies. The FMM method, however, distinguishes itself from other forensic measures by building on earlier work in Klimek et al. (2012) and explicitly modeling the process of stealing and manipulating votes in favor of the incumbent. The FMM procedure then estimates both the extent of incremental (π_2) and extreme fraud (π_3) via Markov Chain Monte Carlo. — *Fraud indicators used:* Estimated shares of extreme fraud (FMM-EXTR), incremental fraud (FMM-INCR) and the sum of both (FMM-SUM).

4 Approach

4.1 Institutional background

In line with the literature and anecdotal evidence, I assume that the decision about whether, where and how much to manipulate the results of national elections is made at the region-level by Russia's 83 governors (Myagkov et al., 2009; Bader and van Ham, 2014; Kobak et al., 2016b; Moser and White, 2017; Kalinin, 2022).[18] Governors' main power lies in controlling the composition of the TECs whereas the members of the RECs are appointed by recommendation of the federal government and PECs are formed ad-hoc close to the election upon nomination by the electorate (OSCE, 2000b, 2004a,b, 2012a,b). Since also all forensic indicators require variation in election outcomes within some higher-level unit of aggregation and PECs are the most granular unit in the election

[18] For a more detailed discussion of governors' motives for manipulation, see Reuter and Robertson (2012).

process, TECs are naturally the most sensible unit of observation to study election fraud in the Russian context. Using the precinct-level data, I thus compute each of the forensic indicators presented above for each district in every election.

4.2 Bayesian Additive Regression Trees

To learn about the relationship between reported fraud and forensic measures in a flexible way without imposing any functional form, I use BART which is an ensemble ML method and is closely related to tree-based methods. Broadly speaking, these methods seek to explain variation in an outcome Y by successively splitting the sample into sub-regions where an explanatory variable X_j is below a threshold s or not. X_j is taken from a pre-defined set of variables $\mathbf{X}$, along with s, chosen to minimize the residual sum of squares (RSS) in Y after the split. Subsequent splits are chosen to minimize the RSS for the sub-samples created in the previous step and so on.

BART improves on this basic approach by creating K trees and using their average prediction to impute Y as is also done in other ML methods such as random forests. Furthermore, similar to boosting, each tree in a BART model is successively updated by randomly perturbing the one from the previous step. These perturbations can encompass different prediction values at the terminal nodes or using more or less complex tree structures. Preference, however, is given to those tree versions which capture variation in Y not yet accounted for by all other trees in the last step. After B iterations of this kind, where the first one is simply the mean of Y, one obtains B average predictions across K trees and the final model is obtained by taking again the average across these. However, since iterations at the beginning are less reliable, one typically drops the first L burn-in samples from B (James et al., 2021). For my analysis, I set the main BART parameters

as $K = 100$, $B = 5,000$ and $L = 50,000$ to assure comparability with Montgomery et al. (2015b).

4.3 Training samples

A fundamental concern for any ML approach is mis-classification which may induce bias in model predictions. A rather common remedy in the ML literature is removing the observations causing this bias from the training data (see the overview by Hort et al., 2023). For my analysis, I thus run the BART model not only on the *Full* sample but also on three restricted samples which each seek to isolate parts of the data less affected by under-reporting and thus more informative about the true relationship between forensic indicators and electoral manipulation.[19]

The first strategy tackles reporting bias by exploiting the fact that many submitted reports did not fall under the definition of fraud in this paper but oftentimes only procedural irregularities, such as illegal campaigning. The *Reporting* sample includes only district-election cells where at least one report of any sort had been filed and should thus remove a large number of false negatives due to lack of knowledge or repression from the estimation. The second approach addresses the possibility that the correlation between fraud reports and forensic indicators could be driven by unobservable area characteristics. Since the 2011 and 2012 elections were only four months apart, such features were likely constant between these two ballots. Akin to a fixed-effects regression model, one can thus eliminate the influence of constant area characteristics by including only *Switchers*

[19] My data and methodology do not allow applying an automated, data-driven procedure to identify influential mis-classified observations like the one proposed by Verma et al. (2021). Pratola et al. (2023) propose methods to identify such influential observations in BART models with continuous outcomes, whereas my analysis uses categorical ones.

Table 2: Balance across samples

	Total			Training data			Test data		
Sample	All	Fraud=1	Fraud=0	All	Fraud=1	Fraud=0	All	Fraud=1	Fraud=0
Full	5,462	950	4,512	4,369	776	3,593	1,093	174	919
Reporting	1,370	950	420	1,096	781	315	274	169	105
Switcher	872	436	436	698	349	349	174	87	87
Reporting Switcher	218	109	109	174	87	87	44	22	22

in fraud reporting status between those two elections in the training data. These switches could be driven by different incentives for fraud, which are typically lower in presidential elections, as well as turnover among governors or changes in their rigging strategies. Since some of switching could also be driven by false negatives in the 2011 election, when GOLOS was less known, I also use a third approach which combines the previous two strategies. This *Reporting Switcher* sample only includes districts which filed any kind of report in both elections but only once reported election fraud.

Table 2 shows the distribution of the main outcome variable *Fraud* in the four samples which are randomly divided each into a 80% training and a 20% test set. The *Full* sample covers 2,731 TECs for the elections 2011 and 2012 with more than 80% not reporting any fraud. Looking at the *Reporting* sample, total observations drop by about 75% and those without fraud report by 90%. About two thirds of this sample documented at least one instance of manipulation. In the *Switcher* sample, the number of observations shrinks further to 16% of the initial data. By construction, the outcome is perfectly balanced in this sample. The same also applies to the *Reporting Switcher* sample which features only 4% of the original data.[20]

[20] For the two *Switcher* samples, I choose test and training sets by randomly selecting among districts rather than district-election cells to ensure that district pairs are not separated in the randomization process which would reduce the small sample size even further.

4.4 Balancing

Table 2 highlights another potential issue: the outcome imbalance for the *Full* and *Reporting* sample. Since ML methods like BART are designed to explain as much of the variation in the data as possible, the predictions from models trained on such skewed datasets will be systematically biased towards correctly predicting the majority class and significantly underperform for the minority class (Chawla et al., 2002). The ML literature has suggested various techniques to overcome this issue, which typically involve over-sampling the minority or under-sampling the majority class. For the purpose of this article, I rely on a widely used technique called SMOTE for the creation of the training data. SMOTE combines under-sampling with an over-sampling technique which creates perturbed synthetic replications of individual observations rather than identical copies thereof (Fernandez et al., 2018).[21] As shown in Chawla et al. (2002), this approach outperforms mere replication of minority class data in terms of predictive performance.

To decide on the optimal degree of SMOTEing, I proceed as follows: first, for a given over-sampling rate of N for the minority class, I under-sample the majority class so that perfect balance across both groups is automatically assured. Then, the BART model is run on the augmented training dataset and calculate its predictive performance for the 20% test data set aside before applying SMOTE.[22] Next, I compare the predictive performance for over-sampling degrees from 0% (the original data) up to 1000% in increments of 100% (with $k = 5$ following Chawla et al. (2002)). Finally, I choose the final degree of over-sampling samples based on the N which maximizes the average predictive performance

[21] Synthetic data points of the minority class are created by randomly interpolating the other variables' values between an observation and a random subset of its k nearest neighbors.

[22] Note that, as the degree of over-sampling the minority class becomes larger, this may actually also imply over-sampling the majority group.

across the two outcome groups to avoid favoring any group due to outcome imbalance in the non-SMOTEd test data. For the two *Switcher* samples, which are balanced by construction, I use the original data without applying SMOTE.

Importantly, this procedure seeks to maximize the BART models' internal validity and implicitly assumes that the samples are not affected by any of the issues discussed in Section 2.3. This makes external validation through secondary data a particularly important part of my analysis.

4.5 Aggregation

As described above, I use data for the elections 2011 and 2012 to train individual BART models which seek to capture the TEC-level relationship between reported fraud and forensic indicators. After estimation, assuming that the uncovered relationships remain constant over time, each model can be used to predict fraud for all elections in my sample. The predicted values from the model are by default continuous in the interval of 0 and 1 and transformed into binary form if they cross some threshold value. For my calculations, I use a threshold of 0.5 in line with common practice in the literature. Analogous to the original outcome variable, one can interpret the resulting values as the predicted occurrence of fraud for the respective sample.

These district-level predictions have two shortcomings. First, the secondary data on electoral integrity presented in Section 2.4 is only available for a cross-section of regions and as country-level time-series. Hence, there is no natural data source to benchmark the estimates against to judge their performance and to evaluate the different approaches to handling mis-classification. The second issue is that the predictions can only credibly

provide information on the occurrence of fraud rather than its intensity like the GOLOS reports used as input variable.[23]

A simple solution to both of the above problems would be to aggregate the TEC-level predictions to higher administrative units. This could be done, for instance, by averaging the binary fraud predictions for each region or the entire country. On the one hand, this would bring the estimates to a level of aggregation where they can be compared with secondary data sources. On the other hand, this variable can be regarded at least as an imperfect proxy of fraud intensity since it is equivalent to the share of districts affected by fraud and provides information about how widespread manipulation was in a particular election. This approach, however, does not consider how many votes came from a particular area and the share of electorate represented by it. Using weighted averages offers a way to take these margins of heterogeneity into account. The resulting measure provides an even closer proxy of fraud intensity as it can be interpreted as an upper bound of the estimated shares of votes or the electorate affected by manipulation.

Since it is ex-ante unclear whether votes or electorate size are the more suitable weights, I calculate higher-level aggregates of the binary predictions not only for each of the four samples but also using as weights the number of votes as well as the electorate size. In total, I thus obtain 8 different aggregate fraud measures. I calculate these for the region- and the country-level and benchmark them by evaluating their correlational strength with region-level ratings of electoral integrity and the country-level manipulation proxy developed by Montgomery et al. (2015b).

[23] This makes them also less valuable for further studies into the drivers of election fraud since they implicitly assume that the perpetrators have no control over the extent of manipulation.

Table 3: Predictive power across BART models

Sample	Over-sampling	Test data: Observations: All	Fraud	No Fraud	% Predictions Correct: All	Fraud	No Fraud	Average	Full data: % Predicted Fraud: 2011	2012
Full	100	1,093	174	919	73.65	63.22	75.63	69.42	29.16	33.17
Reporting	300	274	169	105	64.23	73.37	49.52	61.45	57.19	54.23
Switcher	None	174	87	87	54.02	56.32	51.72	54.02	39.77	56.54
Reporting Switcher	None	44	22	22	65.91	72.73	59.09	65.91	53.79	50.60

5 Results

5.1 Internal validity

I start by evaluating the predictive power and internal validity of the four different BART models. Table 3 shows for each sample the chosen optimal degree of SMOTE over-sampling, the composition of the test data and the percentage of correct predictions.[24] The last two columns display the predicted share of fraudulent districts for the election covered by GOLOS. For these final predictions, I re-train the BART models on the full data, including the 20% test data. The geographical distribution of these four different binary district-level fraud predictions is displayed in Figure 2.[25]

According to Table 3, the *Full* sample achieves the highest specificity with almost 76% correctly classified no-fraud cases while detecting more than 63% of the fraud observations. Unsurprisingly, the corresponding maps in Figure 2 also show a strong resemblance with the original fraud data in Figure 1. While this shows a high degree of internal validity, these predictions are at odds with reality and illustrate how mis-classified training data can tilt the model towards predicting the propensity to report fraud rather than its actual

[24] See Appendix Section A.1 for the detailed SMOTE results.

[25] Appendix Figure C.3 shows the corresponding maps for the continuous, non-binarized prediction values.

occurrence. Even though both elections were widely denounced for their high degree of manipulation, the estimates suggest this affected only about 30% of districts and took place predominantly in the two largest cities and their surroundings as well as smaller, urban districts in other regions.[26]

When using the *Reporting* sample, BART's ability to predict the fraud cases in the test data increases to about 73%. For the no-fraud class, instead, predictive power plummets to 50%. One very likely explanation for this finding is that conditioning on reporting status discards many false negatives but not all. Hence, even though the model does a good job at detecting manipulation, it also correctly assigns the fraud category to many incorrectly classified no-fraud observations. In line with this reading, also the geographical distribution shown in the two corresponding maps seems more credible in the sense that fraud was more pervasive overall, not restricted to urban areas and slightly more concentrated in the Southwest where many ethnic republics are located. Curiously, manipulation is also predicted to be less common in the two largest cities compared to the *Full* sample. Lastly, predicted manipulation levels are higher in 2011 as suggested by the NELDA fraud measure (1.75 in 2011, 1.32 in 2012) and qualitative assessments in the OSCE observer reports (OSCE, 2012b,a).

The *Switcher* sample fares worse for both groups (56% and 52%, respectively). Looking at the data reveals the share of districts with predicted fraud to be about 40% in 2011 and 57% in 2012 which is at odds with the conclusions from secondary data mentioned above. This is most likely due to the disproportionate inclusion of mis-classified no-fraud cases

[26] For the 2011 elections, the executive summary of the OSCE report notes "frequent procedural violations and instances of apparent manipulation, including several serious indications of ballot box stuffing" (OSCE, 2012b), whereas for 2012 it states that "voting was assessed positively overall; however, procedural irregularities were observed" (OSCE, 2012a).

A. Prediction *Full* 2011

B. Prediction *Full* 2012

C. Prediction *Reporting* 2011

D. Prediction *Reporting* 2012

E. Prediction *Switcher* 2011

F. Prediction *Switcher* 2012

G. Prediction *Reporting Switcher* 2011

H. Prediction *Reporting Switcher* 2012

FIGURE 2: PREDICTED ELECTION FRAUD 2011/2012 RESULTING FROM DIFFERENT TRAINING SAMPLES

Notes: Maps of the Russian Federation showing binary fraud predictions for the elections 2011 and 2012 across regions (thick lines) and districts (thin lines). White areas with thick black borders denote missing data.

from the 2011 elections when GOLOS was still less widely known.[27] On the one hand, this makes the comparison between fraud and no-fraud instances less informative and limits the potential for the BART model to learn. On the other hand, this prevents the BART model from detecting fraud types that were more frequently used in 2011 compared to 2012. The *Reporting Switcher* sample, in turn, achieves a sensitivity of about 73% and a specificity of 59% which are the best results among the three restricted samples. The corresponding predictions also yield the highest prevalence of fraud across the four samples with about 54% in 2011 and 51% in 2012 and show a strong concentration of predicted fraud in the Southwest.

These results are better than they may seem at first. The country-level BART model by Montgomery et al. (2015a), for instance, which also featured several contextual variables was able to predict 61% of fraud cases correctly. On the other hand, their model achieved a specificity of almost 94% compared to a maximum of 59% across the three restricted samples. Later, I turn towards secondary data to understand whether this is due to remaining instances of mis-classification in the outcome variable or low explanatory power of the forensic indicators and to accurately benchmark the models' performance.

Lastly, one may worry that the BART estimates are predominantly driven by data from Russia's two largest cities, Moscow and St. Petersburg, which make up about 25% of the fraud-reporting observations. Reassuringly, I find a high correlation between the original estimates for 2011 and 2012 and alternative ones after excluding those cities from the training process of the four samples.[28]

[27] 20% of districts submitted any type of report (fraud or no fraud) in 2011, compared to 29% in 2012.

[28] The correlations are 0.7087 (*Full* sample), 0.5545 (*Reporting*), 0.8252 (*Switcher*) and 0.6777 (*Reporting Switcher*). These correlations are even higher after aggregating to the region-level.

Table 4: Performance across BART models

Cross-section: Correlation of avg. fraud intensity 2007/08 with expert ratings of regions' electoral corruption 2006-10

Weights/Sample	Full	Reporting	Switcher	Reporting Switcher
Votes	−0.2203	0.4926	0.1617	**0.5102**
Electorate	−0.2082	0.4800	0.1558	0.5043

Time-series: Correlation of aggregate fraud intensity with NELDA-based fraud measure at the country-level

Weights/Sample	Full	Reporting	Switcher	Reporting Switcher
Votes	−0.4193	0.4961	0.4424	**0.7587**
Electorate	−0.3667	0.3708	0.3941	0.7125

5.2 External validity

In this part of the analysis, I evaluate the congruence of the fraud intensity measures resulting from the 8 different combinations of sample and aggregation weights with secondary data introduced in Section 4.5. The top panel in Table 4 shows the correlation of the mean estimated fraud intensity in the elections 2007 and 2008 aggregated by region for each configuration with expert assessments of a regions' electoral corruption.

The first thing to note is that fraud intensity based on the *Full* sample shows a negative correlation with electoral corruption. Again, a very likely explanation for this counter-intuitive finding is that, without accounting for reporting bias, the BART model predicted the propensity to report which was highest in progressive areas where people were either informed about GOLOS or not afraid to report fraud. In the *Reporting* and *Switcher* samples, the correlations are positive throughout. The fact that those correlations are consistently stronger in the former, suggests that reporting bias is more problematic than time-invariant unobservables. The *Reporting Switcher* sample performs even better and provides for both aggregation configurations the highest correlation with expert ratings of electoral manipulation. Compared to the impact of the training sample, aggregating by votes or electorate size seems to matter very little in general. The highest correlation

coefficient with a value of 0.5102 is obtained for the *Reporting Switcher* sample when aggregating the estimates via the number of votes cast in each TEC.

The second panel in Table 4 assesses the correspondence of the 8 fraud intensity measures with aggregate country-level trends in manipulation from 2000 to 2018. Consequently, I also aggregate my TEC-level fraud estimates for the entire country instead of the region. The time-series correlations have the expected direction for all fraud estimates and are, on average, stronger than at the cross-sectional level. This is particularly reassuring since changes in the usage of particular manipulation techniques could have rendered the BART predictions less relevant over time. Again, the *Reporting Switcher* sample emerges as strongest in both setups and attains the highest correlation overall for the aggregates using the number of votes with a value of 0.7587. In light of these results, I proceed with the *Reporting Switcher* sample as my baseline measure of fraud intensity for the remainder of the paper.

5.3 Fraud estimates across regions and elections

The benchmarking exercise in Section 5.2 established a high degree of congruence of the predicted fraud levels with secondary data. However, this data was only available at the national level over time and for regions during the time period 2006–2010. In the following, I discuss the changes in predicted manipulation according to the baseline measure from Section 5.2 across regions and elections for the entire sample period. Absent any quantitative benchmarks for this time period, I compare them with qualitative assessments from official OSCE election observer reports and, where necessary, by other sources.

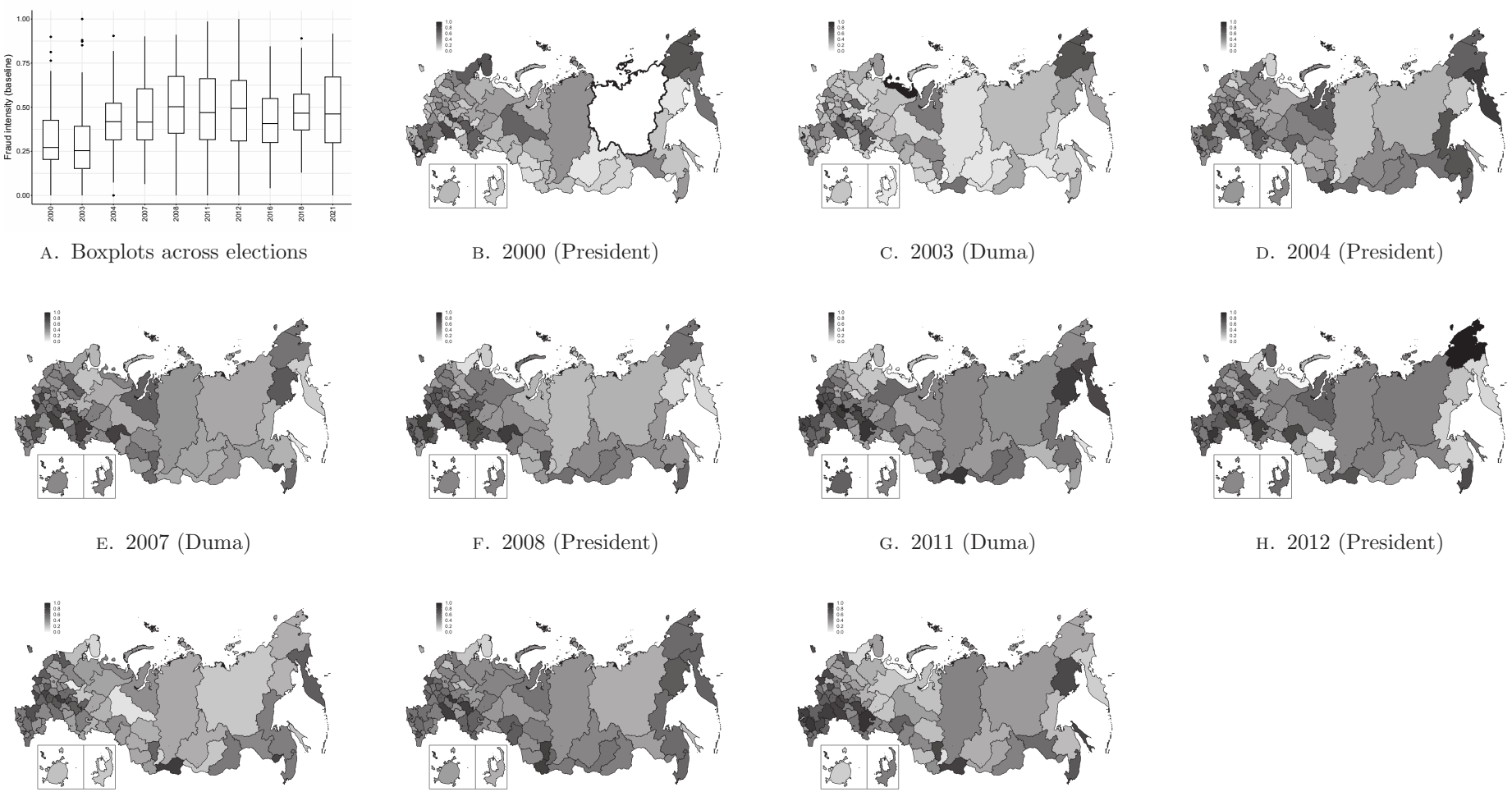
The first graph in Figure 3 shows a series of boxplots which provide a quick overview of the distributional changes in predicted election fraud over time. One can immediately see that predicted fraud levels have notably increased from a mean of 32% in 2000 to 49% in 2008. In terms of the median, fraud estimates nearly doubled during this time period from 27% to 50%. After this initial surge, manipulation broadly remained stable until 2021 with a slight dip in 2016. The remaining graphs in Figure 3 show the spatial distribution of the election-specific boxplots. Looking at the first map for 2000, when Vladimir Putin was elected president for the first time, we can see that most regions on this map retain a rather light color, with low levels of predicted manipulation inspite of some high-intensity regions located in the Southwest. While the OSCE regarded these elections overall as "consistent with international democratic standards", it also acknowledges complaints about manipulation filed by the largest opposition party in the aftermath of the voting. The complaint mentions several regions which also score among the highest in my fraud predictions: Kabardino-Balkariya (90%), Mordoviya (81%), Saratov (76%), Bashkortostan (63%), Tatarstan (67%) and Karachayevo-Cherkessiya (52%) (Belin, 2000; OSCE, 2000a).

In 2003, the mean intensity slightly drops to 30% but fraud remains high in the Southwestern area as well as Chukotka in the Northeast. The OSCE report highlights issues with campaigning, but found that the elections were "generally well-administered" and irregularities in the vote counting process "appeared to be motivated by a desire to speed up the process" (OSCE, 2004a). The report, however, also explicitly discusses instances of carousel voting in the region of Bashkortostan which ranks as one of the highest in terms of estimated fraud intensity in 2003 with a value of 65%. Only three months later, in the presidential elections of 2004, manipulation levels across the country rise substan-

tially to an average of 42%. This deterioration is also reflected in the OSCE report which now states that "the process overall did not adequately reflect principles necessary for a healthy democratic election" and notes cases of falsification and "unauthorized persons apparently directing the work of polling stations" (OSCE, 2004b). The OSCE also explicitly names several regions with implausible, likely fraudulent results which also rank high on my intensity measure: Mordovia (90%), Kabardino-Balkaria (82%), Tatarstan (68%), Karachayevo-Cherkessia (52%), North Ossetia-Alania (48%) and Ingushetia (42%).

For the Duma elections 2007 and the presidential elections 2008 won by Dmitry Medvedev, the average level of predicted fraud rises further to 45% and 50%, respectively. From the maps, one can see that fraudulent (darker) areas start to spread across the country and further intensify in the Southeast, which is in line with observations made in Lukinova et al. (2011). Unfortunately, there was no OSCE mission for 2007 and 2008 due to observers' visa denials and excessive restrictions by the authorities (OSCE, 2007, 2008). The congruence of these estimates with secondary data, however, has been established by their correlation with the expert ratings from Petrov and Titkov (2013) in Section 5.2.

In 2011, the mean fraud estimates remain high at 49%. This mimicks also the increasingly negative assessment by the 2011 OSCE mission which, while not listing particular regions, generally notes "frequent procedural violations and instances of apparent manipulation, including several serious indications of ballot box stuffing" (OSCE, 2012b). The corresponding map shows that, in spite of a constant average, manipulation polarized in the Western part whereas in Central Russia and the Far East several regions flipped their status from high to low intensity and vice versa. For 2012, which marked the return of Vladimir Putin as president, a similar picture emerges with a slightly lower average



A. Boxplots across elections

B. 2000 (President)

C. 2003 (Duma)

D. 2004 (President)

E. 2007 (Duma)

F. 2008 (President)

G. 2011 (Duma)

H. 2012 (President)

I. 2016 (Duma)

J. 2018 (President)

K. 2021 (Duma)

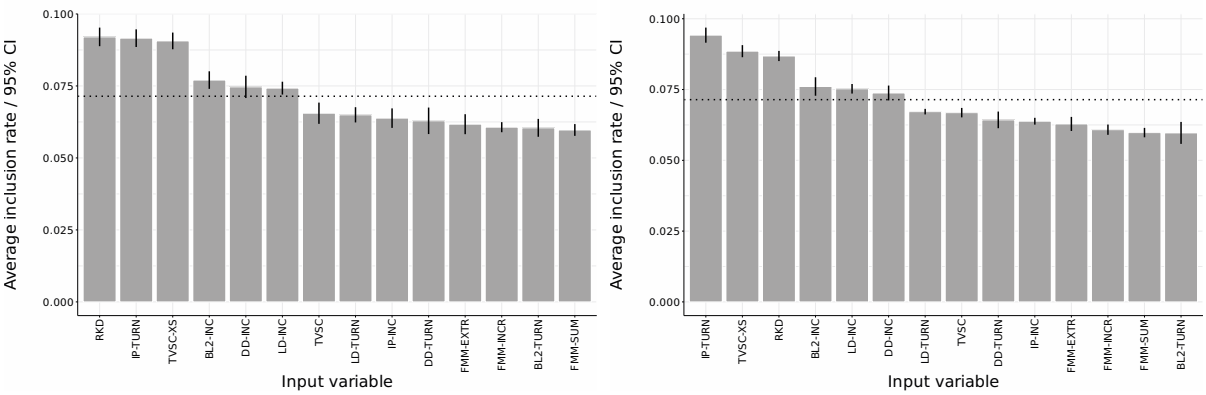
FIGURE 3: ESTIMATED FRAUD INTENSITY ACROSS REGIONS AND ELECTIONS 2000–2021

Notes: Maps of the Russian Federation showing fraud intensity for the elections 2000 to 2021 across regions. White areas with thick black borders denote missing data. The boxplots in the first subgraph show the distribution of fraud intensity over time displayed in the election-specific maps.

intensity of 48%. Also the OSCE arrives at a slightly more positive assessment but still notes "procedural irregularities observed" (OSCE, 2012a).

The discrepancy in manipulation levels between parliamentary and presidential elections observed in 2007/2008 also shows up in the 2016/2018 cycle where my estimates indicate an average of 42% and 47%, respectively. 2016 shows a general reduction in fraud intensity, particular in Western and Central Russia and a slight increase in the Far East. In 2018, fraud levels increased again and became slightly more uniform across the country, as shown in the boxplot. Both OSCE reports, once again, describe observed instances of ballot stuffing (OSCE, 2016, 2018). For 2016, the OSCE also mentions questionable and cancelled results in several regions with above average levels of predicted fraud: Mordovia (84%), Belgorod (74%), Rostov (71%), Nizhny Novgorod (57%) and Saratov (54%).

The most recent national ballot, the Duma election of 2021, continues the high levels of manipulation observed in previous elections with an average of 48%. The spatial distribution of manipulation is also once more strongly concentrated in the Southwest of the country as well as the border regions in the Southeast. While the OSCE again did not send a mission due to excessive restrictions on the number of observers by the Russian authorities, several pieces of evidence support the idea that fraud levels remained high during this election (OSCE, 2021). First, similar to the elections 2011, 2012 and 2016, with comparably high levels of predicted manipulation, there were public protests against the apparent manipulation of results (Devitt and Neely, 2021). Second, as noted by Hutcheson (2022), the decline in citizens' trust in the fairness of elections continued after the 2021 ballot.



A. 10 trees B. 100 trees (baseline)

FIGURE 4: VARIABLE INCLUSION RATES

6 Performance of forensic indicators

6.1 Variable importance

The BART model provides a straightforward way to assess the importance of all input variables which offers a unique opportunity to compare the predictive performance of the various forensic indicators for the Russian context. To do this, I calculate the share of splitting rules in the baseline BART model which include a particular forensic measure. Intuitively, if all variables mattered to the same extent, their inclusion rates should be equal. Hence, if an input variable is used to split the data more often than others, it also has comparatively more predictive power. As shown by Chipman et al. (2010), however, BART tends to include all variables to a similar extent in the splitting rules with a large number of regression trees K, regardless of their actual relevance. To caution against this concern, I calculate inclusion rates for an alternative BART model with ten trees ($K = 10$) as done in Montgomery et al. (2015b) along with the equivalent for the baseline specification with $K = 100$.

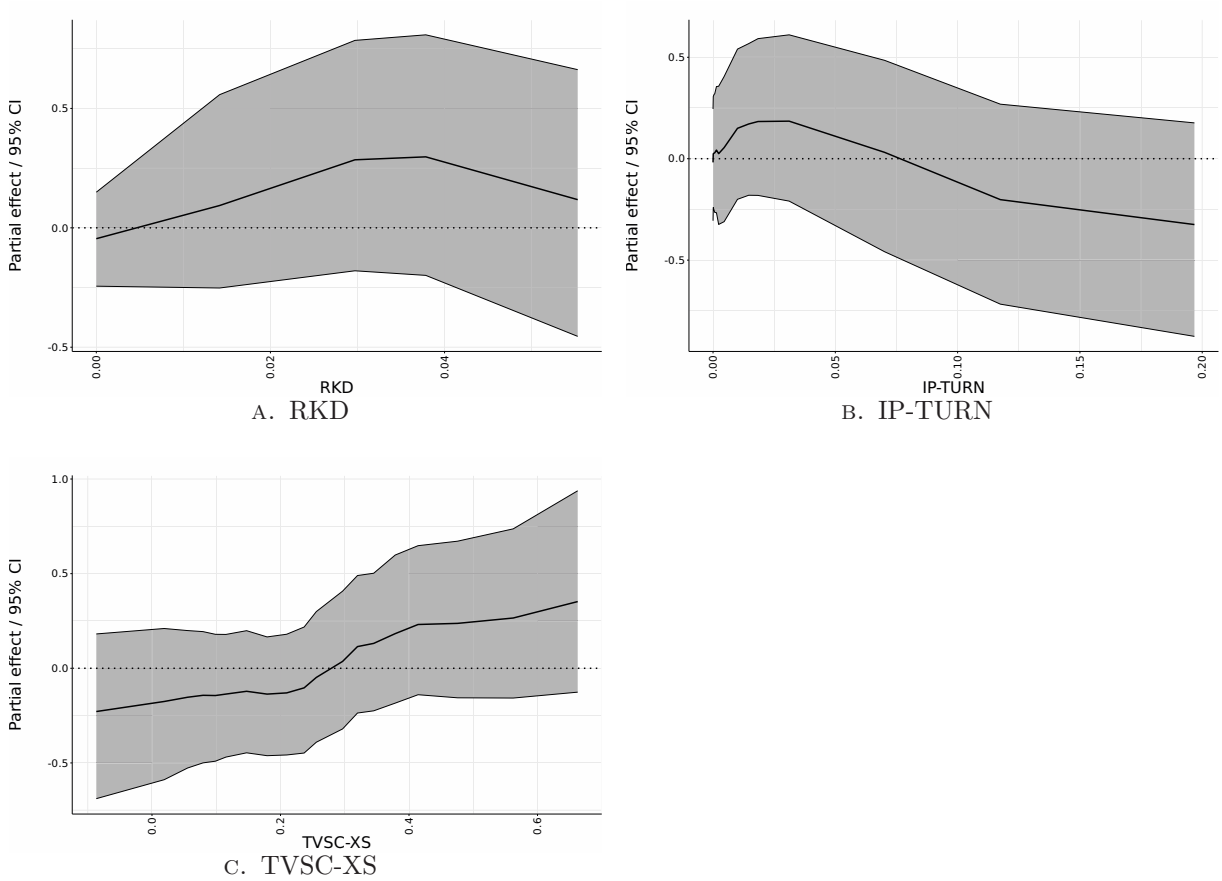
Figure 4 shows all forensic indicators, ordered by their inclusion rate, for both specifications. The dotted, horizontal lines indicate the average inclusion rate 1/14 if all variables

mattered the same. Comparing the two figures, one can see that the number of trees does not strongly affect the results and ordering of variables. In both specifications, the indicators with clearly above-average inclusion rates are RKD, IP-TURN and TVSC-XS. This suggests that during the 2011/2012 elections, results were in fact not manipulated by one single method throughout the entire country but, instead, in some areas through means such as carousel-voting or ballot stuffing whereas in others the votes were likely invented or tweaked to meet particular vote shares and turnout rates.

While the three indicators mentioned above are the best performers, the evidence seems to suggest that also the indicators close to and below the dotted line still hold some explanatory value. The lowest values are attained by FMM-SUM for $K = 10$ and BL2-TURN for $K = 100$ with 0.060 each. If these forensic measures were completely dominated by the ones above the dotted line, one would expect inclusion rates much farther away from the average. In sum, while RKD, IP-TURN and TVSC-XS outperform the other variables in terms of their inclusion rates, the results do not indicate that the other forensic measures can be generally dismissed.

6.2 Effect plausibility

The previous section has shown that some variables correlate more strongly with reported instances of election fraud than others. The precise nature of this relationship, however, could not be revealed in the variable inclusion plots. From a theoretical perspective, the direction of the effects are unequivocal. The RKD and FMM variables are designed to measure shares of fraud among precincts and votes, respectively, and should thus increase with reports of manipulation. This holds also true for TVSC and TVSC-XS, which are supposed to increase in the degree of ballot stuffing or other turnout-inflating types of



A. RKD B. IP-TURN C. TVSC-XS

FIGURE 5: PARTIAL DEPENDENCE PLOTS OF THE MAIN EXPLANATORY VARIABLES' EFFECT ON REPORTED FRAUD

manipulation. All test-based forensic measures consist of a p-value for rejecting the null hypothesis of no abnormal digit or percentage distributions. Particularly low values are thus indicative of manipulation.

Given these clear theoretical expectations, understanding whether they are matched by the empirical relationship estimated in the BART model constitutes a crucial step in assessing the plausibility of a particular indicator in the studied context. For BART models we can investigate these relationships through partial dependence (PD) plots. PD plots show, for a chosen set of quantiles of an explanatory variable, the mean predicted outcome across all trees in the estimated model while fixing all other variables at their actual values. They also feature the corresponding 95% confidence intervals which, given the low amount of observations in the *Reporting Switcher* sample ($N = 218$), are in-

significant throughout. Figure 5 shows the PD plots of the baseline BART model for the three variables which emerged as particularly relevant in Section 6.1: RKD, IP-TURN and TVSC-XS.[29] The graphs display the partial effect by ventiles between the 5th and 95th percentile of the respective variable to avoid distortions from outliers.

Looking at the top left graph in Figure 5 shows the relationship between RKD and reported fraud. The higher the share of fraud according to RKD, the more positive the correlation the correlation with reported fraud. This effect decreases by more than half for RKD values above 0.038 but remains positive. The top right graph shows the partial effect of IP-TURN on reported fraud which, as a p-value, should be positively correlated with fraud for very low values close to zero and exhibit a negative relationship overall. The results are mixed in that respect. While the effect of IP-TURN on reported fraud is indeed negative for values above 0.07, there is virtually no effect on fraud close to zero.

For TVSC-XS in the bottom left plot, the direction of the effect on fraud is fully in line with theory in the sense that it monotonically increases for higher values of the forensic measure. The effect only starts turning positive for TVSC-XS values above 0.30 which, given that also Myagkov et al. (2009) described this case as more ambiguous, do not constitute a clear violation of theory.

Taken together, RKD and TVSC-XS emerge as fairly plausible forensic indicators linked to election fraud in the context of Russia 2000–2021. This does not generally refute IP-TURN nor any of the other indicators as all have been proven to be powerful forensic tools in other studies and empirical settings. In addition, one needs to bear in mind the small sample size and that the entire assessment rests on the assumption that the BART model results were not systematically distorted in any way.

[29] The PD plots for the remaining forensic measures are displayed in Appendix Figure C.4.

7 Conclusion

This article describes a context-independent, data-driven procedure to estimate fraud intensity at the sub-national level via forensic methods. Using the case of post-2000 Russia, I show how crowd-sourced reports of election fraud can be combined with ML methods to train a BART model which is able to predict manipulation in setups where only forensic indicators derived from election micro-data are available. Bearing in mind that the fraud report data is likely to contain many mis-classified cases, I use four different training samples and benchmark these against aggregate secondary data at the cross-sectional and time-series level. My preferred predictions also match well with qualitative accounts of regional election fraud between 2000 and 2021. Given the increasing supply of election micro-data and the spread of CSEM, my approach could potentially be applied in other empirical contexts with limited data availability. Furthermore, my findings highlight a so far overlooked positive aspect of CSEM: the detection of election fraud on a broader scale. International donors may want to consider this insight when deciding whether to support similar initiatives in the future.

References

Ajao, Toyin. 2022. "Ushahidi's Nonviolent Technological Impact in Kenya's 2008 Post-Election Violence". In Akin Iwilade and T. M. Ebiede (Eds.), Youth and Non-Violence in Africa's Fragile Contexts, pp. 163–188. Cham: Palgrave Macmillan.

Alvarez, R. Michael, Thad E. Hall, and Susan D. Hyde (Eds.)2008. Election fraud: Detecting and deterring electoral manipulation. Brookings Series on Election Administration and Reform. Washington and D.C: Brookings Institution Press.

Arias, Carlos R., Jorge Garcia, and Alejandro Corpeño. 2015. "Population as Auditor of an Election Process in Honduras: The Case of the VotoSocial Crowdsourcing Platform". Policy & Internet 7 (2): 185–202.

Bader, Max. 2013. "Do new voting technologies prevent fraud? Evidence from Russia". USENIX Journal of Election Technology and Systems 2 (1): 1–8.

Bader, Max and Hans Schmeets. 2013. "Does International Election Observation Deter and Detect Fraud? Evidence from Russia". Representation 49 (4): 501–514.

Bader, Max and Carolien van Ham. 2014. "What explains regional variation in election fraud? Evidence from Russia: a research note". Post-Soviet Affairs 31 (6): 514–528.

Bailard, Catie Snow and Steven Livingston. 2014. "Crowdsourcing Accountability in a Nigerian Election". Journal of Information Technology & Politics 11 (4): 349–367.

Beber, B. and A. Scacco. 2012. "What the Numbers Say: A Digit-Based Test for Election Fraud". Political Analysis 20 (2): 211–234.

Belin, Laura. 2000. "Fraud Charges Unlikely to Be Examined Impartially". RFE/RL Russian Election Report 13 (5). 7 April, Available at: https://web.archive.org/web/20000815063955/http://www.rferl.org/elections/russia00report/.

Bott, Maja, Björn-Sören Gigler, and Gregor Young. 2014. "The Role of Crowdsourcing for Better Governance in Fragile State Contexts". In B.-S. Gigler, S. Bailur, M. Bott, and G. Young (Eds.), Closing the Feedback Loop: Can Technology Bridge the Accountability Gap?, pp. 107–148. Washington DC: World Bank Publications.

Callen, Michael and James D. Long. 2015. "Institutional Corruption and Election Fraud: Evidence from a Field Experiment in Afghanistan". American Economic Review 105 (1): 354–381.

Cantú, Francisco. 2019. "The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election". American Political Science Review 113 (03): 710–726.

Cantú, Francisco and Sebastián M. Saiegh. 2011. "Fraudulent Democracy? An Analysis of Argentina's Infamous Decade Using Supervised Machine Learning". Political Analysis 19 (4): 409–433.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research 16 : 321–357.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian additive regression trees". The Annals of Applied Statistics 4 (1).

Cochran, William G. 1954. "Some Methods for Strengthening the Common Chi-squared Tests". Biometrics 10 (4): 417.

Deckert, J., M. Myagkov, and P. C. Ordeshook. 2011. "Benford's Law and the Detection of Election Fraud". Political Analysis 19 (3): 245–268.

Devitt, Polina and Jason Neely. 2021. "Hundreds of Russians join Moscow protest over parliamentary election". Reuters. 25 September, Available at: https://www.reuters.com/world/europe/hundreds-russians-join-moscow-protest-over-parliamentary-election-2021-09-25/.

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

Fernandez, Alberto, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. 2018. "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary". Journal of Artificial Intelligence Research 61 : 863–905.

Gandhi, Jennifer and Ellen Lust-Okar. 2009. "Elections Under Authoritarianism". Annual Review of Political Science 12 (1): 403–422.

Grömping, Max. 2017. "Domestic Election Monitoring and Advocacy: An Emerging Research Agenda". Nordic Journal of Human Rights 35 (4): 407–423.

Gunawan, F. and Y. Ruldeviyani. 2020. "Improving Data Quality in Crowdsourced Data for Indonesian Election Monitor: A Case Study in KawalPilpres". Journal of Physics: Conference Series 1566 (1): 012095.

Hellström, Johan. 2015. "Crowdsourcing as a Tool for Political Participation? The Case of Ugandawatch". International Journal of Public Information Systems 11 (1): 1–19.

Herron, Erik S. and Fredrik M. Sjoberg. 2016. "The Impact of 'Boss' Candidates and Local Political Machines on Elections in Ukraine". Europe-Asia Studies 68 (6): 985–1002.

Hort, Max, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey". mimeo. Available at: https://arxiv.org/abs/2207.07068.

Hutcheson, Derek S. 2022. "National Elections in Russia". In G. Gill (Ed.), Routledge Handbook of Russian Politics and Society, pp. 111–126. London: Routledge.

Hyde, Susan D. and Nikolay Marinov. 2012. "Which Elections Can Be Lost?". Political Analysis 20 (2): 191–210.

Hyde, Susan D. and Nikolay Marinov. 2021. "Codebook for National Elections Across Democracy and Autocracy Dataset, 6.0". mimeo. Available at: https://nelda.co/{#}access.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. An Introduction to Statistical Learning: With Applications in R (2 ed.). New York: Springer.

Kalinin, Kirill. 2022. "Signaling Games of Election Fraud: A Case of Russia". Russian Politics 7 (2): 210–236.

Kelley, Judith G. 2012. Monitoring Democracy: When International Election Observation Works, and Why It Often Fails. Princeton and Oxford: Princeton University Press.

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

Kobak, Dmitry. 2023. "Statistical anomalies in Russian elections: Data". Available at: https://github.com/dkobak/elections.

Kobak, Dmitry, Sergey Shpilkin, and Maxim S. Pshenichnikov. 2016a. "Integer percentages as electoral falsification fingerprints". The Annals of Applied Statistics 10 (1).

Kobak, Dmitry, Sergey Shpilkin, and Maxim S. Pshenichnikov. 2016b. "Statistical fingerprints of electoral fraud?". Significance 13 (4): 20–23.

Kobak, Dmitry, Sergey Shpilkin, and Maxim S. Pshenichnikov. 2018. "Putin's Peaks: Russian Election Data Revisited". Significance 15 (3): 8–9.

Kobak, Dmitry, Sergey Shpilkin, and Maxim S. Pshenichnikov. 2020. "Suspect Peaks in Russia's "Referendum" Results". Significance 17 (5): 8–9.

Leemann, Lucas and Daniel Bochsler. 2014. "A systematic approach to study electoral fraud". Electoral Studies 35 (3): 33–47.

Lehoucq, Fabrice. 2003. "Electoral Fraud: Causes, Types, and Consequences". Annual Review of Political Science 6 (1): 233–256.

Levin, Ines, Julia Pomares, and R. Michael Alvarez. 2016. "Using Machine Learning Algorithms to Detect Election Fraud". In R. M. Alvarez (Ed.), Computational Social Science, pp. 266–294. Cambridge: Cambridge University Press.

Lukinova, Evgeniya, Mikhail Myagkov, and Peter C. Ordeshook. 2011. "Metastasised Fraud in Russia's 2008 Presidential Election". Europe-Asia Studies 63 (4): 603–621.

Mack, Verena and Lukas F. Stoetzer. 2019. "Election fraud, digit tests and how humans fabricate vote counts - An experimental approach". Electoral Studies 58 (2): 31–47.

Mares, Isabela and Lauren Young. 2016. "Buying, Expropriating, and Stealing Votes". Annual Review of Political Science 19 (1): 267–288.

Mebane Jr., Walter R. 2008. "Election Forensics: The Second-Digit Benford's Law Test and Recent American Presidential Elections". In R. M. Alvarez, T. E. Hall, and S. D. Hyde (Eds.), Election fraud, Brookings Series on Election Administration and Reform, pp. 162–181. Washington and D.C: Brookings Institution Press.

Mebane Jr., Walter R. 2013. "Using Vote Counts' Digits to Diagnose Strategies and Frauds: Russia". Presented at the 2013 Annual Meeting of the American Political Science, Chicago, IL, August 29–September 1, 2013. Available at: https://ssrn.com/abstract=2303480.

Mebane Jr., Walter R. 2015. "Can Vote Counts' Digits and Benford's Law Diagnose Elections?". In S. J. Miller (Ed.), Benford's Law: Theory and Applications, pp. 212–222. Princeton and Oxford: Princeton University Press.

Mebane Jr., Walter R., Diogo Ferrari, Kevin McAlister, and Patrick Y. Wu. 2022. "Measuring Election Frauds". mimeo. Available at: http://www.umich.edu/~wmebane/measfrauds.pdf.

Medzihorsky, Juraj. 2015. "Election Fraud: A Latent Class Framework for Digit-Based Tests". Political Analysis 23 (4): 506–517.

Montgomery, Jacob M., Santiago Olivella, Joshua D. Potter, and Brian F. Crisp. 2015a. "An Informed Forensics Approach to Detecting Vote Irregularities". Political Analysis 23 (4): 488–505.

Montgomery, Jacob M., Santiago Olivella, Joshua D. Potter, and Brian F. Crisp. 2015b. "Supplemental Materials for: An Informed Forensics Approach to Detecting Vote Irregularities". Political Analysis 23 (4): 488–505.

Moser, Robert G. and Allison C. White. 2017. "Does electoral fraud spread? The expansion of electoral manipulation in Russia". Post-Soviet Affairs 33 (2): 85–99.

Myagkov, Mikhail and Alexander Sobyanin. 1996. "Irregularities in the 1993 Russian Election Returns". National Council for Soviet and East European Research Working Paper 810-05. Available at: https://www.ucis.pitt.edu/nceeer/1996-810-05-Myagkov.pdf.

Myagkov, Mikhail G., Peter C. Ordeshook, and Dimitri Shakin. 2009. The forensics of election fraud: Russia and Ukraine. Cambridge: Cambridge University Press.

OSCE. 2000a. Russian Federation, Elections to the State Duma, 19 December 1999. OSCE/ODIHR Election Observation Mission Final Report. Warsaw.

OSCE. 2000b. Russian Federation, Presidential Election, 26 March 2000. OSCE/ODIHR Election Observation Mission Final Report. Warsaw.

OSCE. 2004a. Russian Federation, Elections to the State Duma, 7 December 2003. OSCE/ODIHR Election Observation Mission Final Report. Warsaw.

OSCE. 2004b. Russian Federation, Presidential Election, 14 March 2004. OSCE/ODIHR Election Observation Mission Final Report. Warsaw.

OSCE. 2007. "ODIHR unable to observe Russian Duma elections". 16 November, Available at: https://www.osce.org/odihr/elections/49175.

OSCE. 2008. "OSCE/ODIHR regrets that restrictions force cancellation of election observation mission to Russian Federation". 7 February, Available at: https://www.osce.org/odihr/elections/49438.

OSCE. 2012a. Russian Federation, Elections to the State Duma, 4 December 2011. OSCE/ODIHR Election Observation Mission Final Report. Warsaw.

OSCE. 2012b. Russian Federation, Presidential Election, 4 March 2012. OSCE/ODIHR Election Observation Mission Final Report. Warsaw.

OSCE. 2016. Russian Federation, State Duma Elections, 18 September 2016. OSCE/ODIHR Election Observation Mission Final Report. Warsaw.

OSCE. 2018. Russian Federation, Presidential Election, 18 March 2018. OSCE/ODIHR Election Observation Mission Final Report. Warsaw.

OSCE. 2021. "No OSCE observers for Russian parliamentary elections following major limitations". 4 September, Available at: https://www.osce.org/odihr/elections/russia/494488.

Petrov, Nikolay and Alexei Titkov. 2013. Democracy Rating of Russian Regions by the Carnegie Moscow Center: 10 Years in Service. Moscow: Carnegie Moscow Center.

Pratola, M. T., E. I. George, and R. E. McCulloch. 2023. "Influential Observations in Bayesian Regression Tree Models". Journal of Computational and Graphical Statistics: 1–17.

Resin, Johannes. 2023. "A Simple Algorithm for Exact Multinomial Tests". Journal of Computational and Graphical Statistics 32 (2): 539–550.

Reuter, Ora John and Graeme B. Robertson. 2012. "Subnational Appointments in Authoritarian Regimes: Evidence from Russian Gubernatorial Appointments". Journal of Politics 74 (4): 1023–1037.

Rozenas, Arturas. 2017. "Detecting Election Fraud from Irregularities in Vote-Share Distributions". Political Analysis 25 (01): 41–56.

Rueda, Miguel R., Guy Grossman, and Shuning Ge. 2023. "Do More Disaggregated Electoral Results Deter Aggregation Fraud?". MIT Political Science Department Research Paper 2023-6. Available at: https://ssrn.com/abstract=4621504.

Rundlett, Ashlea and Milan W. Svolik. 2016. "Deliver the Vote! Micromotives and Macrobehavior in Electoral Fraud". American Political Science Review 110 (01): 180–197.

Salazar, Oscar and Jorge Soto. 2011. "How to Crowdsource Election Monitoring in 30 Days: the Mexican Experience". In M. Poblet (Ed.), Mobile Technologies for Conflict Management, Volume 2, pp. 55–66. Springer.

Shayo, Deodatus Patrick. 2021. "Doing old things in a new way? Technology and crowdsourced observation in the 2015 Tanzanian election". Scientific African 11 (4): e00661.

Shpilkin, Sergey. 2021. "Russian election data: Datasets for analysis of elections in Russia". Available at: https://old.datahub.io/organization/ru{_}elections.

Simpser, Alberto. 2013. Why Governments and Parties Manipulate Elections: Theory, Practice, and Implications. Cambridge: Cambridge University Press.

Sjoberg, Fredrik M. 2014. "Autocratic adaptation: The strategic use of transparency and the persistence of election fraud". Electoral Studies 33 (2): 233–245.

Skovoroda, Rodion and Tomila Lankina. 2016. "Fabricating Votes for Putin: New Tests of Fraud and Electoral Manipulations from Russia". Post-Soviet Affairs 33 (2): 100–123.

Vardanyan, Gegham. 2013. "Elections, Social Movements and Internet Penetration in Armenia". Caucasus Analytical Digest (53-54): 19–22.

Verma, Sahil, Michael Ernst, and Rene Just. 2021. "Removing Biased Data to Improve Fairness and Accuracy". mimeo. Available at: https://arxiv.org/abs/2102.03054.

Zhang, Mali, R. Michael Alvarez, Ines Levin, and Haroldo V. Ribeiro. 2019. "Election forensics: Using machine learning and synthetic data for possible election anomaly detection". PLOS One 14 (10): e0223950.

Supporting Information for Manuscript

"With a Little Help From the Crowd: Estimating Election Fraud with Forensic Methods"

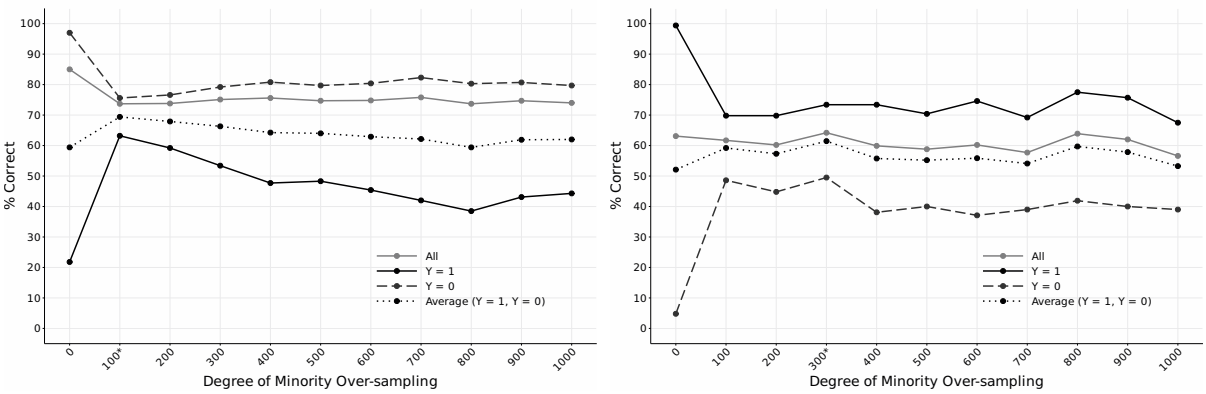
Table of Contents

A Appendix analyses

A.1 SMOTE results

To decide on the optimal degree of SMOTE over-sampling for the *Full* and *Reporting* samples, I evaluate the predictive performance of the BART model for varying degrees of N. The four lines in the plots of Figure A.1 show the shares of correct predictions by over-sampling degree for the full data, the two individual classes with and without any fraud report and the average across the two categories.

The first thing to note is that the bias towards the majority class tackled by SMOTE is indeed an issue in both datasets. In the *Full* sample, where most observations count as no-fraud, using 0% over-sampling results in 97% of all no-fraud instances being correctly classified as opposed to a meagre 22% of the fraud instances. A similar, yet reversed, pattern is observed for the *Reporting* sample, where no-fraud is the minority class. The second observation is that predictive power increases when applying SMOTE in both cases, but that higher levels of over-sampling do not lead to notable improvements. The average accuracy across outcome groups peaks at an over-sampling rate of 100% for the *Full* sample and 300% for the *Reporting* sample.



A. Sample: Full B. Sample: Reporting

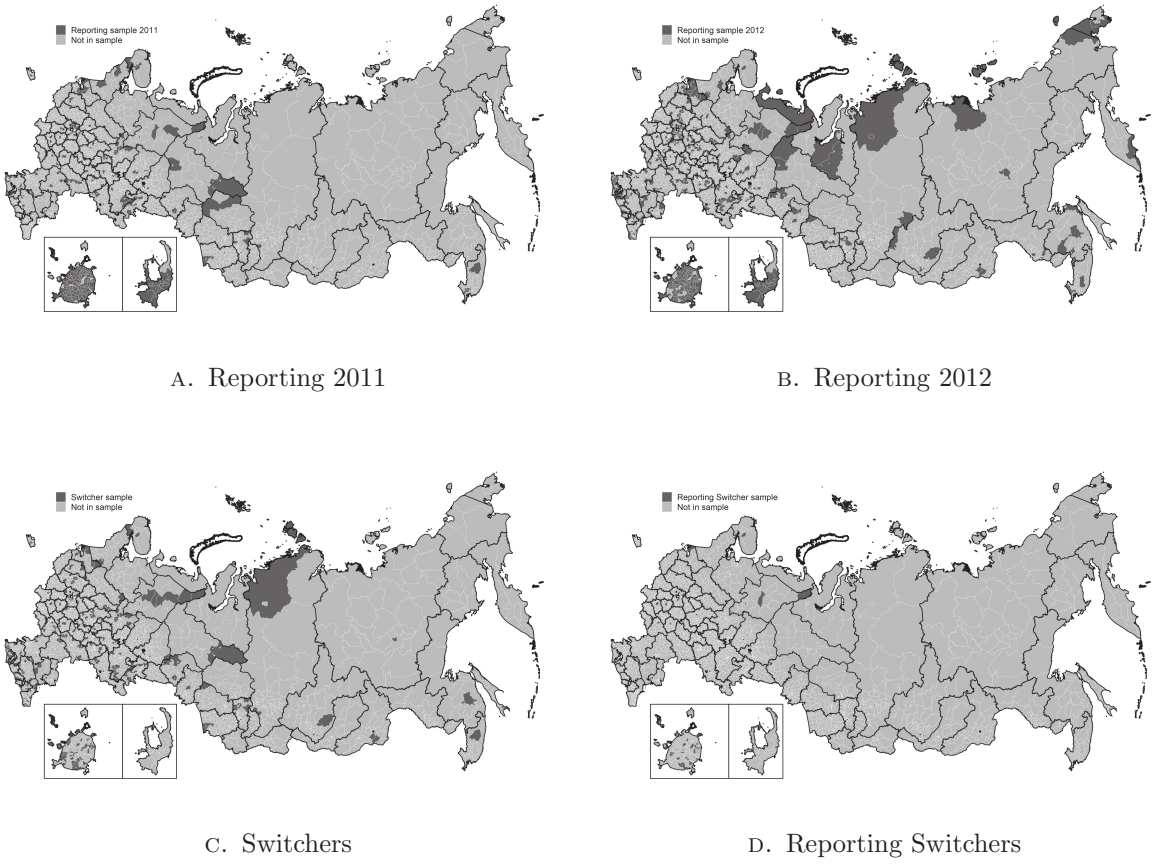
FIGURE A.1: EFFECT OF MINORITY OVER-SAMPLING ON PREDICTIVE PERFORMANCE

B Appendix Tables

Table B.1: Summary statistics

	Obs	Mean	Std.Dev.	Min	Max
Fraud reports (district-level)					
Reported fraud 2011 = 1	2,735	0.16	0.36	0.00	1.00
Reported fraud 2012 = 1	2,735	0.19	0.39	0.00	1.00
Forensic measures (district-level)					
TVSC	30,086	0.83	1.02	−57.98	132.00
TVSC-XS	30,086	0.22	0.99	−58.58	131.28
BL2-INC	30,089	0.76	0.24	0.00	1.00
BL2-TURN	30,089	0.84	0.20	0.00	1.00
LD-INC	30,089	0.51	0.29	0.00	1.00
LD-TURN	30,089	0.52	0.29	0.00	1.00
DD-INC	30,088	0.51	0.29	0.00	1.00
DD-TURN	30,087	0.52	0.29	0.00	1.00
IP-INC	30,204	0.03	0.07	0.00	0.98
IP-TURN	30,204	0.04	0.09	0.00	0.85
RKD	30,099	0.02	0.06	0.00	0.98
FMM-EXTR	30,099	0.07	0.08	0.00	0.48
FMM-INCR	30,099	0.21	0.12	0.00	0.48
FMM-SUM	30,099	0.28	0.16	0.01	0.65
Fraud predictions (district-level)					
Fraud predicted (Full) = 1	30,073	0.29	0.45	0.00	1.00
Fraud predicted (Reporting) = 1	30,073	0.54	0.50	0.00	1.00
Fraud predicted (Switcher) = 1	30,073	0.46	0.50	0.00	1.00
Fraud predicted (Rep. Switcher) = 1	30,073	0.46	0.50	0.00	1.00
Fraud predicted 2011 (Full) = 1	2,726	0.29	0.46	0.00	1.00
Fraud predicted 2011 (Reporting) = 1	2,726	0.57	0.50	0.00	1.00
Fraud predicted 2011 (Switcher) = 1	2,726	0.40	0.49	0.00	1.00
Fraud predicted 2011 (Rep. Switcher) = 1	2,726	0.53	0.50	0.00	1.00
Fraud predicted 2012 (Full) = 1	2,725	0.34	0.47	0.00	1.00
Fraud predicted 2012 (Reporting) = 1	2,725	0.55	0.50	0.00	1.00
Fraud predicted 2012 (Switcher) = 1	2,725	0.57	0.50	0.00	1.00
Fraud predicted 2012 (Rep. Switcher) = 1	2,725	0.51	0.50	0.00	1.00
Fraud intensity and benchmarks (region-level)					
Fraud intensity (Full)	931	0.44	0.20	0.00	0.97
Fraud intensity (Reporting)	931	0.58	0.17	0.00	1.00
Fraud intensity (Switcher)	931	0.46	0.23	0.00	1.00
Fraud intensity (Rep. Switcher)	931	0.43	0.21	0.00	1.00
Electoral corruption rating	89	−3.00	0.94	−5.00	−1.00
Fraud intensity 2007/08 (Full)	83	0.45	0.16	0.00	0.83
Fraud intensity 2007/08 (Reporting)	83	0.57	0.16	0.20	0.95
Fraud intensity 2007/08 (Switcher)	83	0.54	0.13	0.05	0.80
Fraud intensity 2007/08 (Rep. Switcher)	83	0.47	0.17	0.15	0.84
Fraud intensity 2007/08 (Full, Electorate)	83	0.47	0.17	0.00	0.83
Fraud intensity 2007/08 (Reporting, Electorate)	83	0.57	0.16	0.17	0.95
Fraud intensity 2007/08 (Switcher, Electorate)	83	0.54	0.14	0.05	0.80
Fraud intensity 2007/08 (Rep. Switcher, Electorate)	83	0.47	0.17	0.16	0.84
Fraud intensity and benchmarks (country-level)					
Fraud intensity (Full)	10	0.53	0.05	0.45	0.59
Fraud intensity (Reporting)	10	0.61	0.03	0.57	0.68
Fraud intensity (Switcher)	10	0.49	0.16	0.20	0.71
Fraud intensity (Rep. Switcher)	10	0.45	0.09	0.29	0.54
Fraud intensity (Full, Electorate)	10	0.56	0.04	0.49	0.61
Fraud intensity (Reporting, Electorate)	10	0.61	0.03	0.56	0.67
Fraud intensity (Switcher, Electorate)	10	0.47	0.17	0.18	0.72
Fraud intensity (Rep. Switcher, Electorate)	10	0.43	0.08	0.26	0.51
NELDA-based fraud measure	9	1.03	0.69	−0.58	1.75

C Appendix Figures



A. Reporting 2011

B. Reporting 2012

C. Switchers

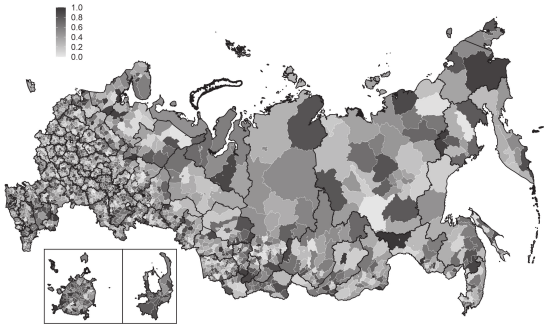
D. Reporting Switchers

FIGURE C.2: RESTRICTED SAMPLES USED IN BART ESTMATION

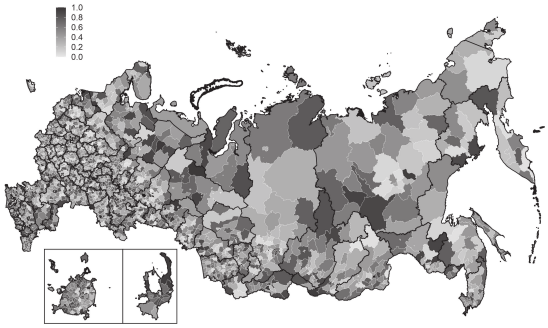
Notes: Maps of the Russian Federation showing the districts belonging to particular restricted samples. White areas with thick black borders denote missing data.

A. Prediction Full 2011

B. Prediction Full 2012



C. Prediction Reporting 2011



D. Prediction Reporting 2012

E. Prediction Switcher 2011

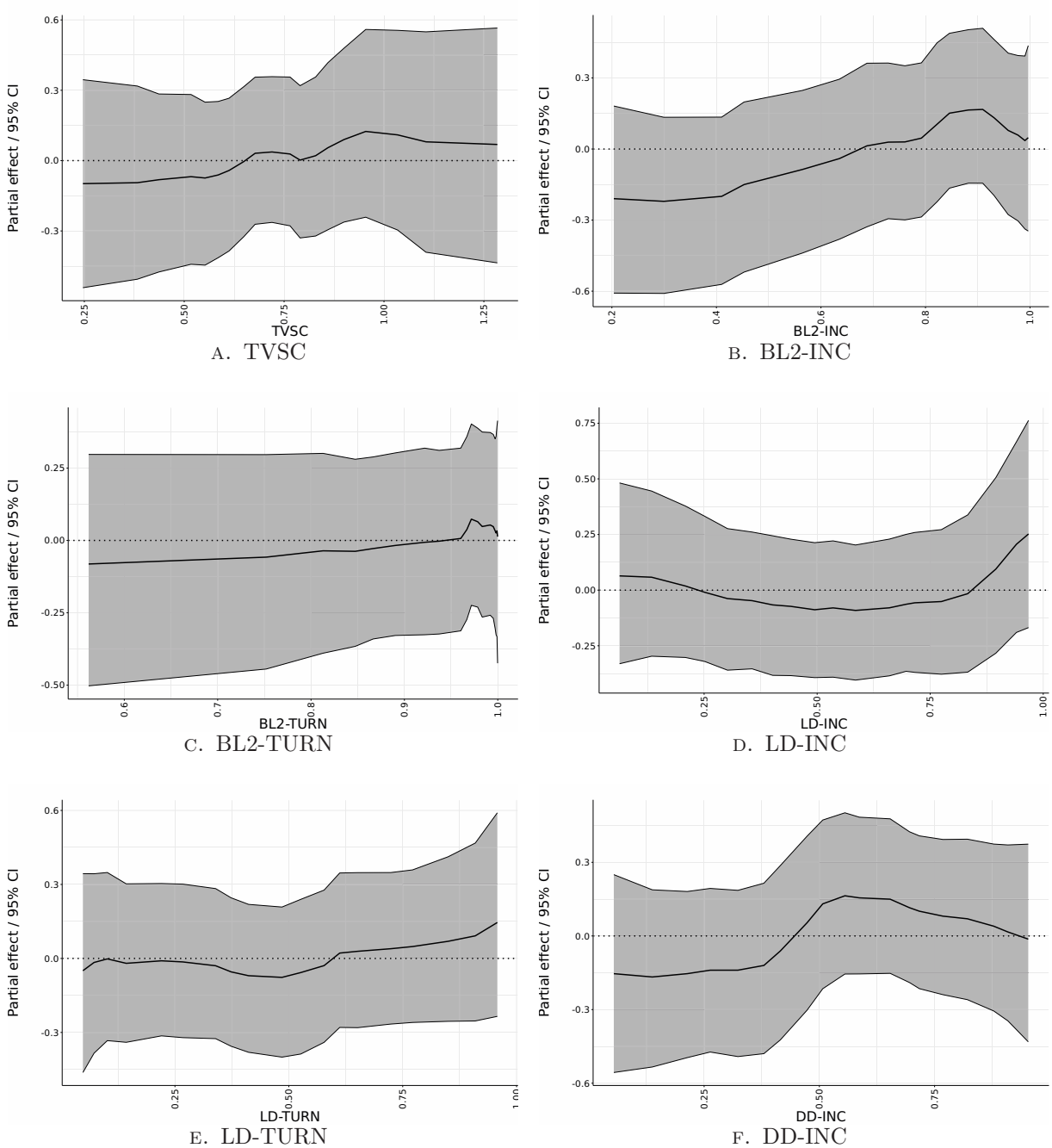
F. Prediction Switcher 2012

G. Prediction Reporting Switcher 2011

H. Prediction Reporting Switcher 2012

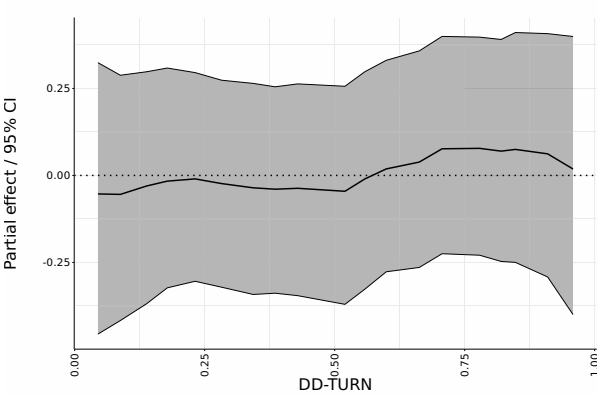
FIGURE C.3: PREDICTED ELECTION FRAUD 2011/2012 RESULTING FROM DIFFERENT TRAINING SAMPLES (CONTINUOUS)

Notes: Maps of the Russian Federation showing continuous fraud predictions for the elections 2011 and 2012 across regions (thick lines) and districts (thin lines). White areas with thick black borders denote missing data.

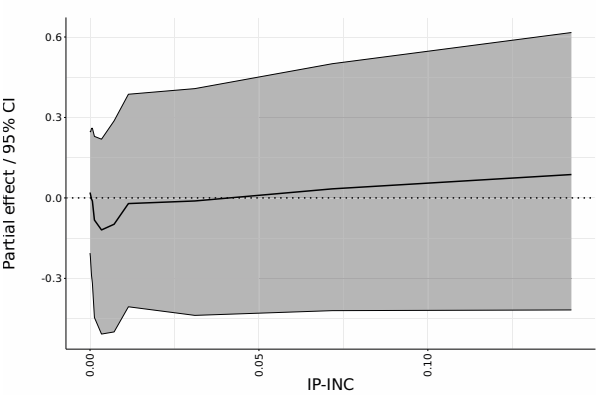


A. TVSC B. BL2-INC C. BL2-TURN D. LD-INC E. LD-TURN F. DD-INC

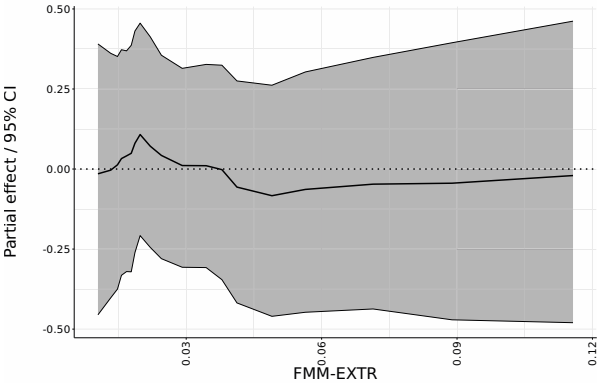
FIGURE C.4: PARTIAL DEPENDENCE PLOTS OF REMAINING EXPLANATORY VARIABLES' EFFECT ON REPORTED FRAUD



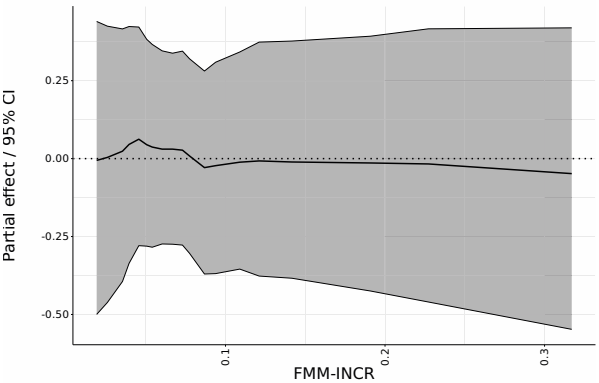
G. DD-TURN



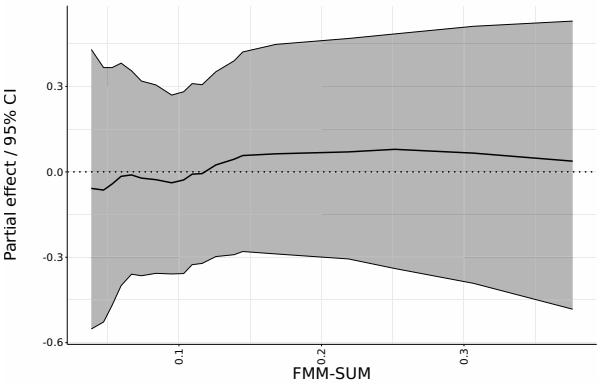
H. IP-INC



I. FMM-EXTR



J. FMM-INCR



K. FMM-SUM

Figure C.4: Partial dependence plots of remaining explanatory variables' effect on reported fraud

RECENT PUBLICATIONS BY *CEIS Tor Vergata*

Biases and Nudges in the Circular Economy: A Review
Luca Congiu, Enrico Botta and Mariangela Zoli
CEIS *Research Paper*, 583 October 2024

Energy Shocks, Pandemics and the Macroeconomy
Luisa Corrado, Stefano Grassi, Aldo Paolillo and Francesco Ravazzolo
CEIS *Research Paper*, 582 August 2024

The Multivariate Fractional Ornstein-Uhlenbeck Process
Ranieri Dugo, Giacomo Giorgio and Paolo Pigato
CEIS *Research Paper*, 581 August 2024

The Macro Neutrality of Exchange-Rate Regimes in the presence of Exporter-Importer Firms
Cosimo Petracchi
CEIS *Research Paper*, 580 July 2024

Monetary Regimes and Real Exchange Rates: Long-Run Evidence at the Product Level
Jason Kim, Marco Mello and Cosimo Petracchi
CEIS *Research Paper*, 579 June 2024

On the Output Effect of Fiscal Consolidation Plans: A Causal Analysis
Lorenzo Carbonari, Alessio Farcomeni, Filippo Maurici and Giovanni Trovato
CEIS *Research Paper*, 578 May 2024

Ordered Correlation Forest
Riccardo Di Francesco
CEIS *Research Paper*, 577 May 2024

Ups and (Draw)Downs
Tommaso Proietti
CEIS *Research Paper*, 576 May 2024

Human Capital-based Growth with Depopulation and Class-size Effects: Theory and Empirics
Alberto Bucci, Lorenzo Carbonari, Giovanni Trovato and Pedro Trivin
CEIS *Research Paper*, 575 April 2024

Optimization of the Generalized Covariance Estimator in Noncausal Processes
Gianluca Cubadda, Francesco Giancaterini, Alain Hecq and Joann Jasiak
CEIS *Research Paper*, 574 April 2024

DISTRIBUTION

Our publications are available online at www.ceistorvergata.it

DISCLAIMER

The opinions expressed in these publications are the authors' alone and therefore do not necessarily reflect the opinions of the supporters, staff, or boards of CEIS Tor Vergata.

COPYRIGHT

MEDIA INQUIRIES AND INFORMATION

For media inquiries, please contact Barbara Piazzi at +39 06 72595652/01 or by e-mail at piazzi@ceis.uniroma2.it. Our web site, www.ceistorvergata.it, contains more information about Center's events, publications, and staff.

DEVELOPMENT AND SUPPORT

For information about contributing to CEIS Tor Vergata, please contact at +39 06 72595601 or by e-mail at segr.ceis@economia.uniroma2.it