An Informed Forensics Approach to Detecting Vote Irregularities*

Jacob M. Montgomery
Assistant Professor of Political Science
Washington University in St. Louis
Campus Box 1063, St. Louis, MO 63130
jacob.montgomery@wustl.edu

Santiago Olivella
Assistant Professor of Political Science
The University of Miami
1300 Campo Sano Avenue, Coral Gables, FL 33146
olivella@miami.edu

Joshua D. Potter
Assistant Professor of Political Science
Louisiana State University
240 Stubbs Hall, Baton Rouge, LA 70803
jpotter@lsu.edu

Brian F. Crisp
Professor of Political Science
Washington University in St. Louis
Campus Box 1063, St. Louis, MO 63130
crisp@wustl.edu

*Replication data and code is available at http://bit.ly/1gMngya. We are grateful for helpful comments we received from Chris Zorn and two anonymous reviewers.

ABSTRACT

Electoral forensics involves examining election results for anomalies in order to efficiently identify patterns indicative of electoral irregularities. However, there is disagreement about which, if any, forensics tool is most effective at identifying fraud, and there is no method for integrating multiple tools. Moreover, forensic efforts have failed to systematically take advantage of country-specific details that might aid in diagnosing fraud. We deploy a Bayesian additive regression trees (BART) model – a machine-learning technique – on a large cross-national dataset to explore the dense network of potential relationships between various forensic indicators of anomalies and electoral fraud risk factors, on the one hand, and the likelihood of fraud, on the other. This approach allows us to arbitrate between the relative importance of different forensic and contextual features for identifying electoral fraud and results in a diagnostic tool that can be relatively easily implemented in cross-national research.

1 INTRODUCTION

Electoral fraud disrupts the chain of responsiveness that links politicians to their supporters and, by so doing, brings to power "representatives" who do not reflect the will of the people. When detected – or even suspected – fraud erodes the legitimacy of the democratic process and can provoke violent unrest, repression, and even civil war. Unfortunately, it remains extremely difficult to detect instances of fraud. Perpetrators of electoral fraud are highly motivated to conceal their acts from opposition parties, the press, and election monitors. As a consequence, confidently determining when fraud has occurred and when it has not is challenging.

In this paper we propose a novel approach for estimating the extent to which an election was likely characterized by fraud that relies only upon subnational election returns and a few widely available pieces of information about the context of the election. Our method builds upon *electoral forensics* techniques, an appealing approach to identifying fraud that has received considerable scholarly attention in recent years. The forensic approach involves analyzing vote results for anomalous numerical patterns. Forensic methods are attractive for at least three reasons. First, they can be applied to election results as soon as they are reported. Second, the anomalous patterns of interest are not specific to the election, country, or culture in question. Finally, the search for anomalous patterns can be applied to any set of results (including historical ones), even if monitors were not present at the election.

Despite their promise, forensic techniques have several limitations. To begin, it is unclear if any single anomaly or pattern can by itself be taken as evidence of fraud. Each forensic tool has its own theoretical basis, with unique strengths and weaknesses. Relying on any one in isolation may be inefficient or, perhaps, even misleading. Nevertheless, surprisingly few

projects have sought to empirically test the relative predictive power of the various forensic tools or sought to combine the various forensic tools in some systematic manner.

Further, while many practitioners of electoral forensics acknowledge that information about case-specific contextual factors should be used to inform the approach, far fewer systematically integrate such knowledge into their analyses. Scholars of electoral fraud who focus on more "substantive" – or non-forensic – analyses of elections have identified several contextual risk factors that increase the likelihood of electoral fraud, including elements as varied as socioeconomic inequalities and district magnitude.Their work clearly suggests that, for instance, elections held in Norway are *a priori* less likely to be fraudulent than those held in Congo and that relatively stronger evidence should be required before declaring Norwegian results fraudulent based on anomalous digit distributions. Ideally, we should augment forensic analyses to systematically incorporate this kind of contextual information without engaging in an *ad hoc* exercise of turning to idiosyncratic, case-specific features.

In this paper, we propose what we call an *informed forensics* approach to identifying fraud in a large cross-national setting. Drawing on the forensics literature, we specify an array of forensic indicators for detecting anomalous returns (e.g., digit distributions) that we include in our model. We augment – or inform – these forensic tools with several widely available, country-specific risk factors that past research suggests increase the probability that fraud will be perpetrated during an election. We create a cross-national dataset spanning 70 countries and six decades containing three sets of variables: (1) an *explicit measure* of likely fraud constructed from evaluations by election monitors and other political actors, (2) *forensic indicators* of anomalous vote distributions, and (3) *contextual risk factors* that, while not directly measuring fraud, have been identified in past research as increasing the likelihood of fraud.

In order to combine forensic indicators and contextual risk factors, we fit a Bayesian additive regression trees (BART) model – a machine-learning technique. By allowing the BART model to explore the dense network of potential relationships between different forensic indicators and contextual risk factors, on the one hand, and the propensity for fraud (as captured by our explicit measure), on the other, this approach is able (1) to improve the out-of-sample predictive performance of the model and (2) to arbitrate between the relative importance of different features – both forensic and contextual – for identifying irregularities.

The paper proceeds as follows. First, we briefly review the literature that applies forensic indicators to electoral returns, focusing on methods that compare electoral returns to theoretic baselines. We also review the non-forensics, more contextual literature on election fraud to identify the risk factors associated with malfeasance. In Section 2.3, we combine these forensic indicators and contextual factors with a novel measure of likely defrauded elec-

tions constructed from the *National Elections Across Democracy and Autocracy (NELDA)* dataset (Hyde and Marinov, 2012). In Section 3, we describe the BART model, which we subsequently set to learn from this cross-national dataset. In Section 4, we show that forensic indicators and context-specific risk factors *in combination* allow us to make better out-of-sample predictions of likely fraudulent elections than relying on either approach alone, that the variables most dispositive for identifying elections that were likely defrauded are a combination of contextual risk factors and forensic tool, and that we can validate our informed forensic approach by comparing the model's predictions with alternative measures of electoral malfeasance.

2 FORENSIC TOOLS AND CONTEXTUAL RISK FACTORS

Forensic methods seek to use recorded votes to find anomalies consistent with human manipulation (Mebane, 2008). The literature contains a rich array of methods, with most works falling into two categories: (1) those that compare results to a *theoretical* baseline and (2) those that compare results to some *empirical* baseline. We briefly review only tools from the theoretical baseline approach here because they are superior for our purposes. Tools that rely on an empirical baseline (e.g., comparing results across geographic areas or comparing current returns to historical results) tend to be more case-specific, making them less appropriate given our goal of building a method for detecting fraud that can be quickly applied in any country and at any point in time.

2.1. *Forensic methods for identifying fraud*

Forensic indicators based on *theoretical* baseline distributions take observed distributions of specific integers as they appear in aggregated election results and compare them to digit distributions that arise when results are generated "naturally."[2] These indicators make use of predictions about how often specific integers or combinations of integers should appear in naturally occurring datasets.

Perhaps the most quintessential forensic indicators rely on theoretical baselines derived from *Benford's Law* (Benford, 1938). While it may seem intuitive that each integer should appear as the first significant digit equally often, this intuition is incorrect. Instead, under specific conditions, the integer one occurs about 30% of the time and each successive integer occurs increasingly less often, with the integer nine occurring less than 5% of the time. Extremely disaggregated, low-level counts (where results are inherently capped) or vote outcomes where an effort has been made to create districts with roughly equal numbers of

[2] This means that the numbers are neither assigned nor influenced by human goals. They also cannot be governed by exogenously imposed minima or maxima.

voters may not conform because they are not entirely natural.[3] Fortunately, as Mebane (2010) notes, while low-level vote counts rarely have first digits that satisfy Benford's Law, there is no reason why they should not have *second digits* that do. Contrary to expectations about first and second digits, Mebane and Kalinin (2009) reason that, unlike first or second significant digits, each integer should occur equally often in the last digit of reported turnout figures. Thus, we examine both the distribution of integers in the *final digit* of election returns as well as deviations from Benford's Law in the *second digit*.

Beber and Scacco (2012) show that individuals fabricating numbers exhibit a preference for pairs of adjacent digits suggesting that such pairs (e.g., 23) should be overly-abundant in fraudulent vote tallies. Likewise, subjects avoid pairs of distant numerals, suggesting that distant pairs (e.g., 28) should appear with lower frequency. Further, Beber and Scacco (2012) also show that individuals underestimate the likelihood of integer repetition in numerical sequences, even when incentivized to create truly random data. This implies that they should observe (relatively) few instances of repeated integers (e.g., 44) in manipulated vote tallies. Our measure of distance between integers will capture this phenomenon (a distance of zero) – but only for the last pair of digits. Based on this experimental work then, unusually low or unusually high average distances across the final two digits may indicate anomalies.

Finally, several scholars examine the first moment (i.e., the mean) of the second significant digit distribution (Grendar, Judge and Schechter, 2007; Cantú and Saiegh, 2011; Mebane, 2012). They contend that this allows for the detection of variations away from "Benford-like" distributions rather than assuming some specific expected distribution. For our purposes here, the work of Cantú and Saiegh (2011) is particularly relevant. Their approach is to learn from synthetic data created to resemble two sets of historic, district-level election results from the province of Buenos Aires, Argentina (one from an election widely acknowledged to have been rigged and another from one widely recognized to have been fair). They then used this synthetic data to train a naive Bayes classifier to recognize instances of fraud. We follow up on their work by including the *first moment of the second significant digit distribution* among the forensic indicators we explore.

Drawing on this very brief review of some of the literature that employs forensic tools based on theoretic baselines, we built a set of forensic indicators for as many country-elections as possible within the constraints of available data. To construct our forensic tools, we collected district-level electoral results from the two of the most prominent repositories of comparative electoral data: the *Global Elections Database* (Brancati, 2007) and, the *Constituency-Level Elections Archive (CLEA)* (Kollman et al., 2011). By combining these datasets, we were able to assemble returns from 598 elections in a diverse set of 70 countries

[3] See Mebane (2010) and Cho and Gaines (2007) on this point.

around the world, with significant coverage in each of the world's major geopolitical regions (see Appendix A for a list of included cases). Importantly, our cases include both places where previous work suggests that elections have been "squeaky clean" and places where outcomes have been viewed with skepticism.

Our data consist entirely of district-level returns from which we calculate O_i as the observed frequency of integer $i \in [0, 9]$ appearing in the second significant digit in party-district vote totals for election i (Mebane, 2012). We can then calculate $\chi^2_{2BL} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$, where E_i is the expected frequency of integer i according to the second-digit Benford's law (2BL). Second, following Grendar, Judge and Schechter (2007), we include the first moment of the second significant digit distributions. We also include a measure of the degree to which the distribution of integers in the *final* digit differs from the expected uniform distribution – $\chi^2_{Uniform}$.

Finally, following Beber and Scacco (2012) we also include measures for the mean distance between integers in the last pair of party-district vote totals (see Appendix B for additional coding information and descriptive statistics).[4]

2.2. *Context-specific risk factors*

Recall that we are trying to build a means of detecting fraudulent election outcomes that combines the insights and methodologies of both the *election forensics* and *context-specific* approaches to studying electoral fraud. As we have argued above, we think that a forensics approach has many positive features for both practitioners and scholars, but the intuition behind *informing* the forensics approach is that the degree of irregularity in election results necessary to justify a "fraudulent" diagnosis should differ across contexts. In order to preserve some of the appealing characteristics of forensic approaches to detecting irregularities, we have focused on risk factors that, first, have predictive power across many (if not all) cases and, second, are readily available for a significant number of cases (see Appendix B for coding decisions).

The scholarly literature focusing on pre-election institutional and socioeconomic factors that increase the danger of fraud being perpetrated later has identified two types of risk factors: variables that create environments where elites will be more motivated to commit

[4]Before moving on, it is worth noting that at least two of these particular forensic tools have traditionally focused on data at a very low level of aggregation (e.g., precinct-level data). The cross-national election results we use below are generally available only at a higher level of aggregation. However, these forensics tools only require that – at some point in the chain of moving from an individual's vote to a national-level outcome – there is someone with the opportunity to intentionally alter the results. In theory, then, the manipulation of election returns can occur at any level of aggregation, whether it be precinct, local, regional, or national level. For instance, some studies have suggested that fraud was most prevalent at the *national* level during the 2004 Venezuelan Presidential election (see, for instance, Martín, 2011). Hence, we believe that it is possible for irregularities to occur at levels of reporting *above* the level of the polling station.

fraud and features that, given some actor's desire to engage in malfeasance, affect the ease or difficulty of effectively manipulating outcomes (Lehoucq, 2003).

In terms of motivating elites, Boix (1999) and Acemoglu and Robinson (2006) argue that great economic disparities cause elites to fear unfavorable electoral outcomes. When the economic stakes of losing the election are high, this produces incentives to win elections at whatever cost. Similarly, ethnic fractionalization may increase the probability that a group will fear turning over power after losing an election (Lehoucq, 2003). Losing control of the government to an ethnically distinct sub-population increases the stakes relative to cases where the population is more homogeneous. Finally, the geographic distribution of voters has also been cited as an indicator of the cost of losing an election (Lehoucq, 2003). In sparsely populated, relatively homogeneous districts, political stakes are not nearly so high as they are in urban areas (Dardé, 1996; Domínguez and McCann, 1996). Based on these works regarding motivation to commit fraud, we include measures of *socioeconomic inequality*, *ethnolinguistic fractionalization*, and the *rural/urban distribution* of the population in our models.

Turning from motivations to opportunities, failure by large swaths of the population to show up to vote has been associated with greater opportunities to manipulate vote totals (Cox and Kousser, 1981; Schedler, 2002). To capture this leeway offered potential perpetrators of fraud, we include a measure of *turnout* in our models. Further, where democratic norms and institutions – broadly conceived – are limited, elections may manifest both autocratic elements and democratic elements at the same time (Levitsky and Way, 2002). In addition, Birch (2007) shows that experience with elections is a primary determinant of the extent to which voters have the ability to estimate likely levels of relative support for parties before elections (Birch, 2007; Kitschelt et al., 1999), with dramatic departures from this "electoral heuristic" serving as a signal that perhaps something is amiss. To capture both of these latter constructs, we create a nominal measure of *regime type* that takes on five possible values based on the "level of background democracy" (Birch, 2007) and democratic age.[5]

Finally, Lehoucq (2003) argues that district magnitude may be related to both the cost of losing an election *and* the opportunity to engage in fraud. He reasons that increasing district magnitude should decrease the motivation to engage in fraud in part because electoral outcomes are less high-stakes as district magnitude (and vote-to-seat proportionality)

[5] This discretization will help make the machine learning process computationally manageable, as described below. Countries receiving a lagged Polity IV score ranging from -10 to -6 are coded as autocracies ($n = 23$). Countries in the middle range of values (-6 to 5) are coded as anocracies ($n = 64$). Democratic countries, countries with Polity scores above 5, were divided into two categories. Old democracies ($n = 422$) had lasted more than ten years, while new democracies ($n = 77$) had ten years or less experience with elections (Beck et al., 2001).

increases. In terms of opportunity, district magnitude tends to be associated with higher levels of turnout (Blais, 2006; Cox, 1999), and high turnout complicates efforts to engage in fraud. We include *average district magnitude* in our tests, reasoning that higher magnitudes may reduce both the motivation to engage in fraud and the ease of carrying it out.

Finally, we control for three potentially important variables that are not mentioned in the context-specific literature on fraud, but nevertheless might systemically contribute to the likelihood of fraudulent activity on behalf of political elites.[6] *Economic crisis* and *regime crises* – such as civil wars, insurgencies, or coups – take into account short-term, within-country dynamics that might prompt incumbent politicians to engage in malfeasance. For the first, we use change in GDP per capita, as reported by the World Bank, in the calendar year before the election as proxy and the indicator for the second was a dichotomous recoding of the Polity IV regime variable which takes a value of 1 in the case of coups, revolutions, state failures, and fractional periods. The final control variable is *election commission independence*, the presence of which we expect to tie the hands of political elites wishing to commit fraud. Using data from the Institute for Democracy and Electoral Assistance on election management, we code as 0 those election commissions managed by the government directly; 1 those commissions that are partially managed by the government; and 2 those commissions that are completely independent of government influence.n

2.3. *A proxy measure of electoral irregularities*

In order for our statistical model to learn which forensic indicators and risk factors are most closely associated with the incidence of fraud, we must first be able to identify cases in the past that were more or less fraudulent on which to train it. Of course, if we had the means to create a large dataset where we knew with *certainty* which elections had been tainted there would be much less controversy in "the real world" over election outcomes and this entire exercise would be unnecessary. However, outside the use of a crystal ball, the means of perfectly classifying elections as defrauded or not does not exist, and we must build a proxy measure for fraud on which to train our model combining forensic tools and context specifics.

The *National Elections Across Democracy and Autocracy (NELDA)* dataset (Hyde and Marinov, 2012) measures a number of good *proxies* for fraud (in contrast to the predictors or risk factors discussed above). We focused on (1) items that co-varied strongly (providing reliability) and (2) items that satisfy the conditional independence assumption of standard measurement models.[7] The measures were then combined using a standard three-parameter

[6] We thank an anonymous reviewer for bringing these additional variables to our attention.

[7] Conditional independence, sometimes termed local independence, holds when two observed indicators are independent of each other conditioned on the common latent trait being measured. This implies that

Table 1. Estimates of fraudulent elections for merged database (n=598)

Fraud score	Frequency	Percentage	Examples
-0.41	524	87.63	Switzerland (1947-2007), Spain (1977-2008), Canada (1945-2008)
0.35	8	1.34	Bangladesh (2001), Philippines (1992,1995,2004), Zambia (2006)
0.49	36	6.02	Colombia (1998, 2006), Italy (1983, 2001, 2006), Venezuela (1973, 1978)
0.57	2	0.33	Romania (2004), Sri Lanka (2001)
0.74	2	0.33	Turkey (1991, 1995)
1.04	5	0.84	Guyana (2001), Jamaica (1967), Mexico (1991), Malawi (1999, 2004)
1.06	1	0.17	Thailand (1992)
1.10	1	0.17	Guyana (1997)
1.19	5	0.84	Kenya (1997), Albania (2005), Pakistan (2002), Sri Lanka (2010)
1.53	4	0.67	Mexico (1994), Pakistan (2008), Philippines (2007, 2010)
1.56	2	0.33	Jamaica (1976, 1980)
1.62	1	0.17	Turkey (1999)
1.64	1	0.17	Dominican Republic (1994)
1.67	2	0.33	Brazil (2002), Singapore (2006)
1.79	2	0.33	Cyprus(1981), Kenya (1992)
1.94	1	0.17	Cameroon (2002)
2.08	1	0.17	Cameroon (1997)

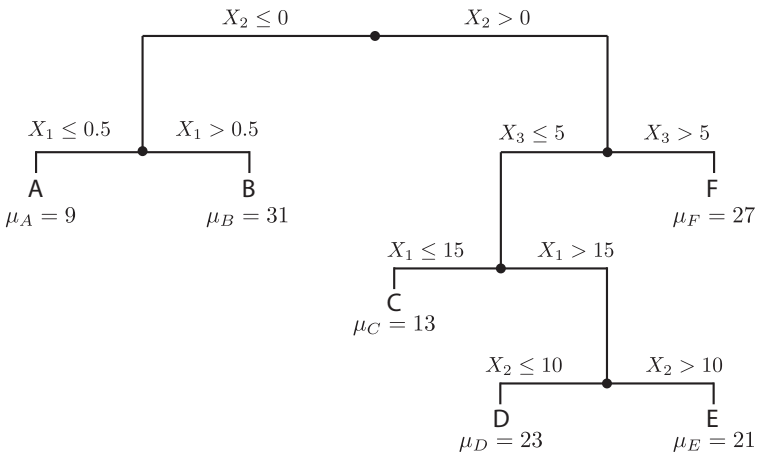
item response theoretic (IRT) model. Additional details about the items used and the measurement model are provided in Appendix B.

Table 1 provides the full distribution of our estimated fraud proxy for the cases that can be matched up to the district-level vote returns needed to calculate forensic indicators, as well as some exemplar cases that fall into each category. Unsurprisingly, allegations of fraud and election irregularities are fairly uncommon in countries willing to provide full district-level election results. Fully 85.8% of all cases were included in our lowest category, providing no overt indication that electoral fraud had likely occurred. For the remaining cases, however, there is quite a bit of variation as to how fraudulent the election appears according to our metric, with the most likely fraudulent legislative elections appearing in Cameroon (1997, 2002), the Dominican Republic (1962), and Equatorial Guinea (1993) – all elections with significant irregularities that have been documented by both academic and journalistic sources.[8]

$p(x_1, x_2|y) = p(x_1|y)p(x_2|y)$, where x_1 and x_2 are indicators and y is the latent trait of interest.

[8] The countries that we include in our training and test sets from NELDA are limited to countries for which district-level election data is readily accessible. The failure to make available district-level returns may disproportionately come from countries with weak electoral infrastructures. These may also be the places, then, where some form of abnormality or some level of malfeasance is most likely. Thus, we are building our informed forensics approach on a set of cases where we are relatively less likely to find abnormalities. This is data limitation that plagues forensics approaches more generally. The included elections, those with district-level returns, do appear to be significantly less fraudulent on average than excluded elections. Although we believe this does not undermine the value of the general approach we describe, we caution against injudicious extrapolation beyond the sample of countries that provide district-level election data.

Figure 1. A example regression tree



3 AN INFORMED FORENSICS APPROACH

The promise of the forensic approach lies in the applicability of the method across widely diverging contexts, requiring neither on-the-ground monitoring, idiosyncratic reliance on case-specific factors, nor special knowledge about a given country or election. Thus far, we have described our cross-national dataset containing several of the most prominent *forensic indicators* of anomalous outcomes, *case-specific risk factors* that provide context in which a country is more susceptible to fraudulent behavior, *and* a proxy for the extent to which an election was likely defrauded. Our goal now is to apply a non-parametric statistical technique from the machine learning literature to that dataset in order to explore how these variables can best be combined to detect irregularities. In Section 4, we then assess the quality of this model. Before turning to this assessment, however, we provide some additional information and intuition about the statistical model.

Bayesian additive regression trees (BART) are a tree-based approach to estimating a flexible functional relationship, $g(.)$, between an outcome variable of interest Y and a set of predictors $\mathbf{X}$ (Chipman, George and McCulloch, 2010; Green and Kern, 2012; Hill, 2012). In our case, the variable of interest (Y) is our proxy fraud score and the set of predictors ($\mathbf{X}$) includes both our forensic tools and our contextual risk factors.

In their simplest form, tree-based approaches approximate the relationship between outcome Y and $\mathbf{X}$ by first partitioning the universe of covariates into k mutually exclusive combinations of covariate profiles and then specifying expected values $\mathbf{M} = (\mu_1, \mu_2, \ldots, \mu_k)$. The partitioning is accomplished by generating a series of inequality-based binary splits of the form $x_n \leq c$ and $x_n \geq c$, where c is some arbitrary threshold. This process can be thought of as the traversal of a binary tree T from its root to one terminal node, at which point an expected value is assigned to that profile.

As an example, the binary tree in Figure 1 partitions the space defined by three covariates (x_1, x_2, x_3) into six mutually exclusive covariate profiles. The terminal nodes are labeled A through F, and the predictions at each terminal node are labeled $(\mu_A, \ldots, \mu_F)$. Thus, terminal node A provides prediction μ_A for all observations such that x_2 is non-positive and x_1 is less than or equal to 0.5. Similarly, a traversal of this tree shows that an observation with $(x_1 = 18, x_2 = 3, x_3 = -4)$ leads to terminal node D, implying $\mu_D = 23$.

In general, given a full specification of a the binary tree T, we can take any observation (i.e. a vector of values on all covariates) and "move down" the tree by following the splitting rules that define the branches until a terminal node is reached. Equipped with a fully specified tree T and a vector of predicted values $\mathbf{M} = (\mu_1, \mu_2, \ldots, \mu_n)$, this assignment can be captured by a function $g(\mathbf{X}, T, \mathbf{M})$, which is a mapping of covariate profiles $\mathbf{X}$ to expected values $\mathbf{M}$ for a given tree T.

To complete the regression tree formulation, we embed this functional relationship in a stochastic framework such that Y is connected to the function $g(.)$ *in expectation*. One general formulation of a regression tree is,

$$Y \sim \mathcal{D}(g(\mathbf{X}, \mathbf{T}, \mathbf{M}), \boldsymbol{\gamma}) \tag{1}$$

where $\mathcal{D}$ is some probability distribution and $\boldsymbol{\gamma}$ are ancillary distribution parameters. By treating T and $\mathbf{M}$ as parameters to be estimated, the regression-tree framework becomes amenable to computation.

Bayesian additive regression trees (or BART) refine this basic regression-tree model in two notable ways (Chipman, George and McCulloch, 2010). First, it builds multiple trees and combines them additively. Specifically, for J trees, the model becomes

$$Y \sim \mathcal{D}\left(\sum_{j=1}^{J} g(\mathbf{X}, T_j, \mathbf{M}_j), \boldsymbol{\gamma}\right) \tag{2}$$

This sum-of-trees approach inherits the flexibility of single-tree models in terms of modeling interactions of various orders, and it adds the possibility of modeling smooth main effects.

Second, the potential for overfitting is reduced by placing regularization priors over the model's parameters (i.e. T, $\mathbf{M}$ and $\boldsymbol{\gamma}$). These priors serve to limit the size and complexity of trees, reducing somewhat the probability of severe over-fitting. Assuming a normal distribution as the likelihood $\mathcal{D}$, a Markov chain Monte Carlo (MCMC) backfitting algorithm can be used to explore the posterior distribution, sampling (at each iteration), a modified set of trees, terminal-node value assignments, and distribution parameters.[9] Additional details

[9] In this case, we need to sample the variance parameter for the assumed normal distribution.

about the BART model are provided in Appendix F.

4 RESULTS

Having reviewed the basic principles of the BART model, in this section we apply the technique to our dataset.[10] Recall that the dataset we constructed includes (1) forensic indicators of anomalous digit distributions calculated from district-level election returns, (2) contextual risk-factors theorized to create an *a priori* context in which fraudulent behavior is ultimately more likely, and (3) our proxy for observed fraud constructed from the NELDA dataset. Our major claim is that electoral fraud, as proxied by our NELDA indicator, will be most easily detected when relying on *both* forensic indicators *and* contextual risk factors together relative to detecting fraud using either set in isolation.

We test this hypothesis in three distinct ways. First, we explore which combination of variables (i.e., contextual, forensic, or both) best *predicts* our electoral fraud proxy out-of-sample by fitting three distinct BART models. As expected, our results suggest that the combined – or *informed forensics* – model performs best. Second, we unpack the role of the individual variables in identifying fraudulent elections within sample. In so doing, we also explore the functional relationships between our variables and the fraud proxy, allowing us to assess whether they are each related to it in the anticipated manner. Third, we conclude this section by validating the BART predictions against two alternative indicators of fraudulent elections in the literature.

4.1. *Comparing BART models*

We randomly divided the observations into training and test sets, comprised of 83.6% ($n = 500$) and 16.4% ($n = 98$) of our data, respectively. The goal is to fit a BART model with the training portion of the data using the *fraud proxy* as our dependent variable, and then validate the model using the remaining observations allocated to the test set.[11] Using

[10] Replication data and code is available at `http://bit.ly/1gMngya`

[11] We selected hyperparameter priors, which control the level of shrinkage, by performing a bootstrapping study where the BART model was fit to 100 bootstrapped samples for each of 32 possible settings of the hyperpriors. Our test-set was excluded entirely from this process, and the bootstrap samples were constructed using the moving block sampling approach discussed below. In total, this involved fitting a total of 3,200 individual BART models. We settled for the parameters that resulted in the highest consensus rank, computed based on the measures of generalization error discussed in Appendix E. The chosen parameters are as follows: `power=0.5`, `base=.95`, `sigdf=10`, `sigquant=0.75`, `ntree=100` and `k=2`. All models were estimated using 5,000 iterations after allowing for a burn-in period of 50,000. In general, this is far more than is needed to allow for convergence, although this was not checked for each and every model.

Table 2. Out-of-sample predictive performance of three BART fits

	RMSE	MAE	MAPE	MAD	MEAPE	Consensus Rank
Informed Forensics	0.407	0.234	0.430	0.090	0.222	1.0
Forensic Tools Only	0.486	0.297	0.549	0.153	0.377	3.0
Contextual Risks Only	0.417	0.238	0.440	0.090	0.223	2.0

$n = 98$

the data that was not employed either in fitting the BART model or in choosing the model hyperparameters (see Footnote 11), we now compare the performance of three BART model fits. We compared the performance of each model using five model fit statistics drawn from the forecasting literature (Brandt, Freeman and Schrodt, 2014): root mean squared error (RMSE), mean absolute error (MAE), mean absolute proportional error (MAPE), median absolute deviation (MAD) and median absolute proportional error (MEAPE). MAPE is measured as a ratio of the error to the value of the dependent variable. Additional details about each fit statistic are shown in Appendix C. For each fit statistic, lower values indicate superior fit.

The results are shown in Table 2.[12] All five metrics show that the *informed forensic* model outperforms a BART model fit using either the forensic and contextual variables in isolation.[13] More specifically, the consensus rank, defined as the mean ranking of the models as evaluated by each fit statistic, shows the informed forensics approach to provide the most-accurate out-of-sample prediction (lower rank indicates superior fit).

One potential concern with these results is that, given that roughly five out of six of our observations take on the lowest fraud score, these fit statistics may simply be capturing the models' accurate prediction of cases with no obvious fraudulent activities. That is, the model has high specificity but low sensitivity. To evaluate whether our model has a very high specificity at the expense of sensitivity, we created binary versions of the outcome variable and the out-of-sample prediction from the BART model. Specifically, we created a bivariate outcome variable that took on a value of 0 for negative scores ($n = 80$, 81.6% in the out-of-sample test set) and 1 for all others. Likewise, we created a variable that took on values of 0 for all cases where the prediction was negative ($n = 82$, 83.7% in the out-of sample test

[12] For each of the three models, we used 50,000 iterations as a burn-in period, and kept the next 5,000 iterations. Standard diagnostics indicated sufficient convergence of the σ BART parameter.

[13] It could also be of interest to evaluate how well BART does when compared to simpler models – such as an off-the-shelf linear model, or a LASSO-regularized linear model – since such models are easier to implement and interpret. The use of BART is justified, however, by the fact that is has consistently lower RMSE values than those simpler models. In particular, an OLS estimation of the informed forensics model yields an RMSE of 0.458, while a LASSO of the same model yields an RMSE of 0.456. The lasso model was fit using the `glmnet` package in `R` using a cross-validated λ parameter (Friedman, Hastie and Tibshirani, 2010).

set) and 1 for all other cases. Of those cases that were high on the fraud score ($n = 18$), the predictions from BART model were also high in 11 instances, giving this binary version of the BART prediction an out-of-sample sensitivity rate of 61.1%. Of those cases that were low on the fraud score ($n = 80$), they were also low on the BART prediction score 75 times, giving the model an out-of-sample specificity rate of 93.75%. In addition, the precision rate of the model is 68.75, indicating that roughly 69% of the cases predicted to be fraudulent by the model actually were. These results suggest that while the model is more accurate in predicting non-fraudulent elections, it is not so to the point where the its value is entirely or even largely based on its ability to predict non-fraudulent election. Indeed, while far from perfect, there seems to be a good balance here between uncovering cases of fraud where they exist while not pointing to an abundance of "false positives."[14]

A further potential objection to these results is that the superior performance of the informed forensics model is somehow a function of the single random partition of the dataset we happened to choose. Ideally, we would like to be able to calculate a single statistic that represents the error rates we would observe across all possible partitions (Efron and Tibshirani, 1997). Unfortunately, this is a difficult task to accomplish due to the sensitivity of the BART algorithm to the composition of the training set. As Hastie, Tibshirani and Friedman (2009) note, "With [...] methods like trees, cross-validation and bootstrap can underestimate the true error [...], because the search for best tree is strongly affected by the validation set. In these situations *only a separate test set will provide an unbiased estimate* of test error" (p. 254, emphasis ours). Thus, the results presented in Table 2 represent the best estimates of the out-of-sample accuracy of the various models, which supports our main claim. That is, cross validation and bootstrap methods designed to infer the error rate of a model independent of a specific training sample are generally inaccurate. In particular, such estimates can be biased, making it far more difficult to distinguish and arbitrate between competing models (Hastie, Tibshirani and Friedman, 2009). However, with this caution in mind, we conducted two additional robustness checks, using well established estimates of *generalization error* (viz. a k-Fold based measure, and a Leave-One-Out Bootstrap based measure), which are out-of-sample test error rates that are less dependent on the specific training sample chosen. We find that the results presented in Table 2 hold, although measures of uncertainty are often too high to make definitive claims about statistical distinctness. For more details on these measures of generalization error, we refer interested readers to Appendix E.

[14]Researchers interested in increasing the sensitivity of the model at the cost of its specificity (and overall model fit) could achieve this goal by choosing hyper-priors for the BART model that are more aggressive in the sense of allowing for significantly less shrinkage. See Chipman, George and McCulloch (2010) for additional discussion of these issues.

4.2. *Which variables matter?*

We next turn to a slightly different question: which particular variables are most important in allowing the fitted BART model to identify electoral fraud (as captured by our NELDA-based proxy)? We expect that the most important factors in the fitted BART model should be a mix of forensic tools and contextual risk factors. To evaluate the importance of each variable in the model, we fit a BART model using the entire ($n = 598$) dataset. When the number of trees is low, BART tends to rely on the most relevant predictors when building trees, as predictors are forced to compete to improve the model's fit (Chipman, George and McCulloch, 2010). As a result, we can re-estimate the BART model using a small number of trees (viz. 10) and calculate the average share of splitting rules that involve each variable.

A variable scoring 0.10 would, on average, be used in one out of every ten binary splitting rules – suggesting that this variable is relatively important for diagnosing fraud. On the other hand, values closer to zero indicate that the BART model largely ignores that variable when constructing the trees.[15] Thus, higher values are obtained by variables that play a more important role in predicting our fraud proxy.

The results, shown in Figure 2, confirm our expectation that the fitted BART models rely on a mix of contextual and forensic factors – although the contextual factors are clearly most critical. The most important forensic indicators are the distance between integers in the final two digits and violations from the uniform distribution in the final digit. The most important contextual factors are the level of ethnolinguistic fractionalization, national turnout rates and the percent of the population living in urban centers.

4.3. *Uncovering relationships*

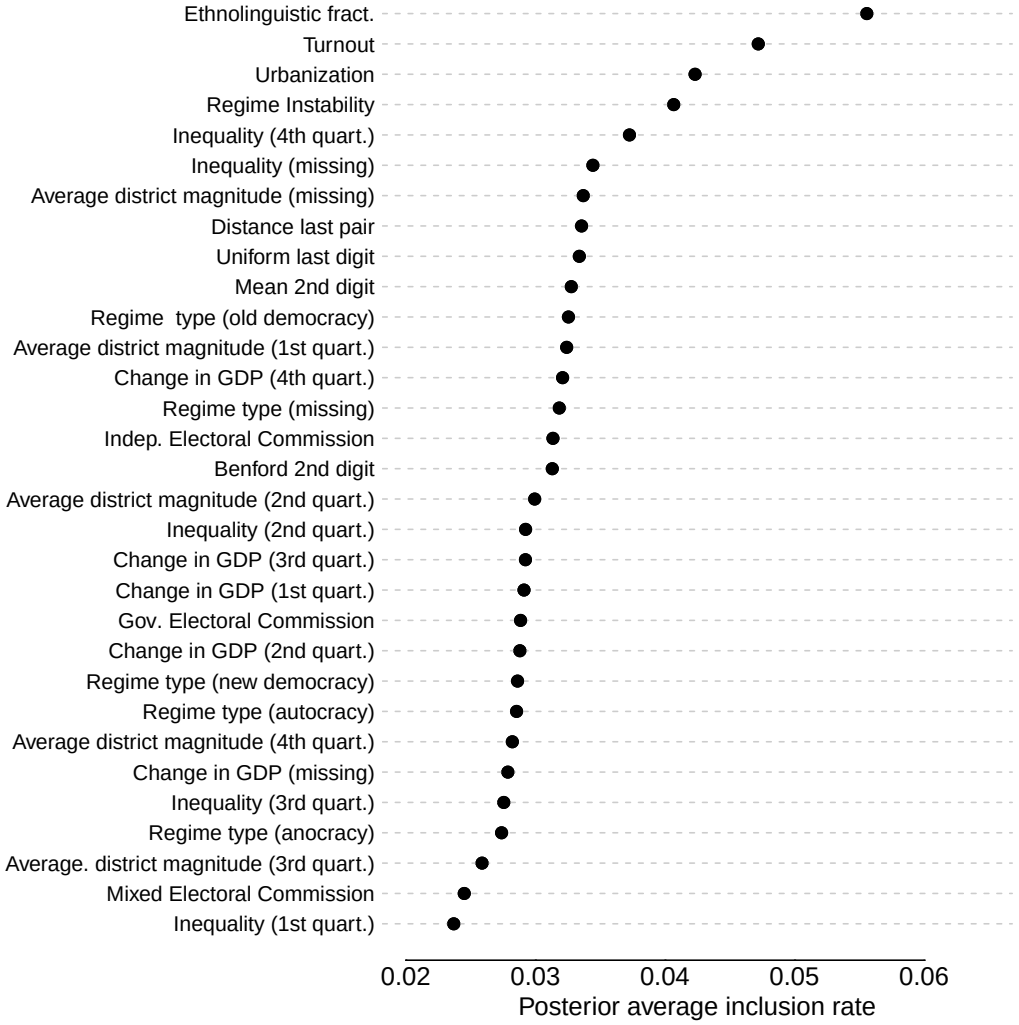
To further unpack the role of the various forensic and contextual factors in the fitted BART model, we display the partial dependence plots for each explanatory covariate by estimating average predicted values of our fraud measure as each variable spans its observed range (or takes on all its possible categorical values) in Figures 3 and 4.[16]

Contextual risk factors: Figures 3 and 4 show that several of the contextual variables *are* independently related to our proxy for electoral fraud. For the most part these relationships conform with the context rich scholarship we reviewed above. For example, political instability (Figure 4, top left panel) has an important effect on the level of fraud, as captured by

[15] Let $I(x_i \in T_j)$ be an indicator function for whether tree j contains covariate x_i. In a model fit with J trees where the posterior is sampled P times, the posterior inclusion probability is $\frac{\sum_{j=1}^{J}\sum_{p=1}^{P} I(x_i \in T_j)}{JP}$.

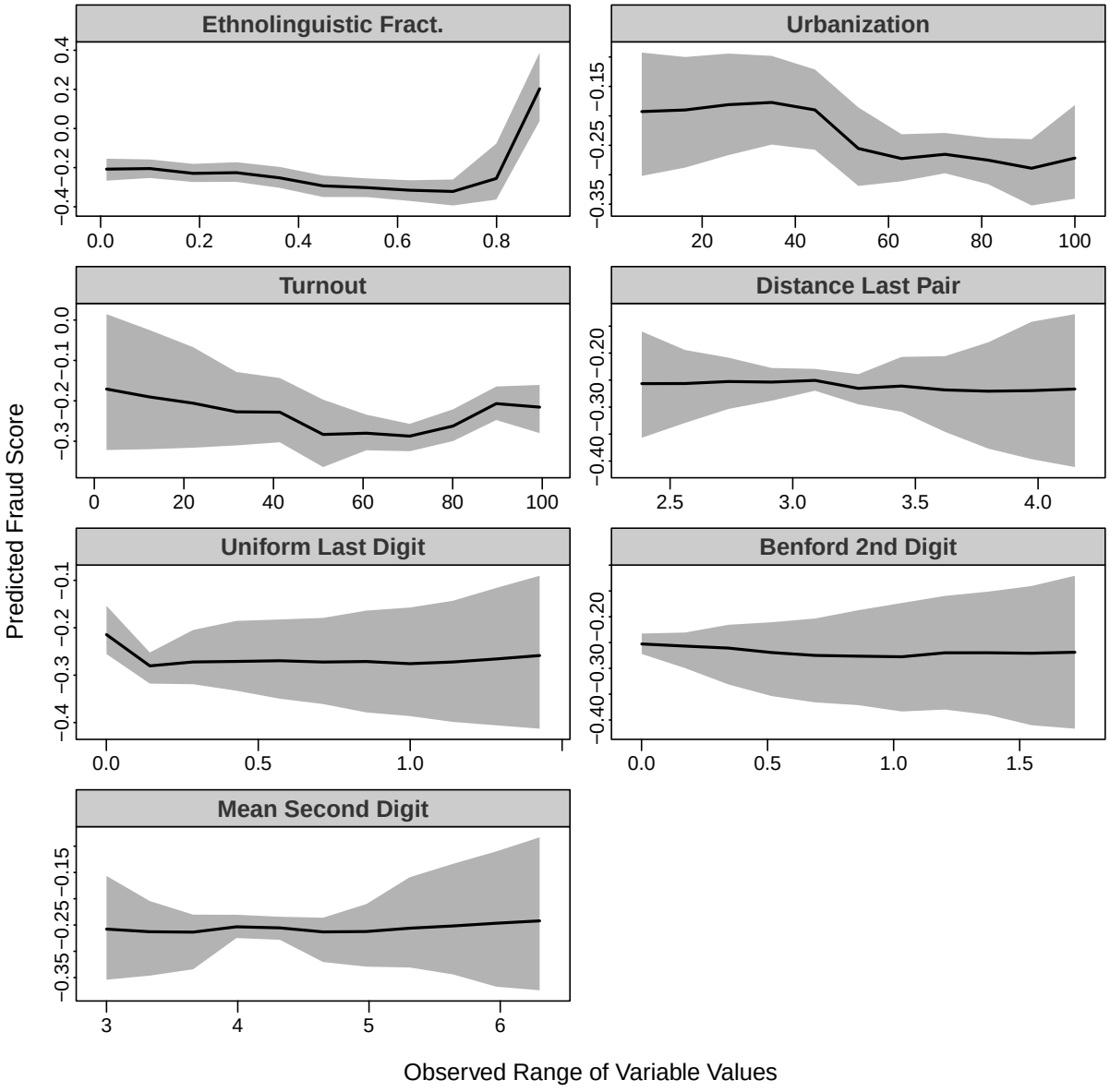
[16] Partial dependence is estimated by making predictions for each observation for different values of the variable in question (holding all other covariates at their actual values) across multiple draws from the posterior. See Chipman, George and McCulloch (2010) for additional description of this procedure. These quantities were estimated using the `pdbart` function in the `BayesTree` package in `R`.

Figure 2. Average inclusion probabilities by variable



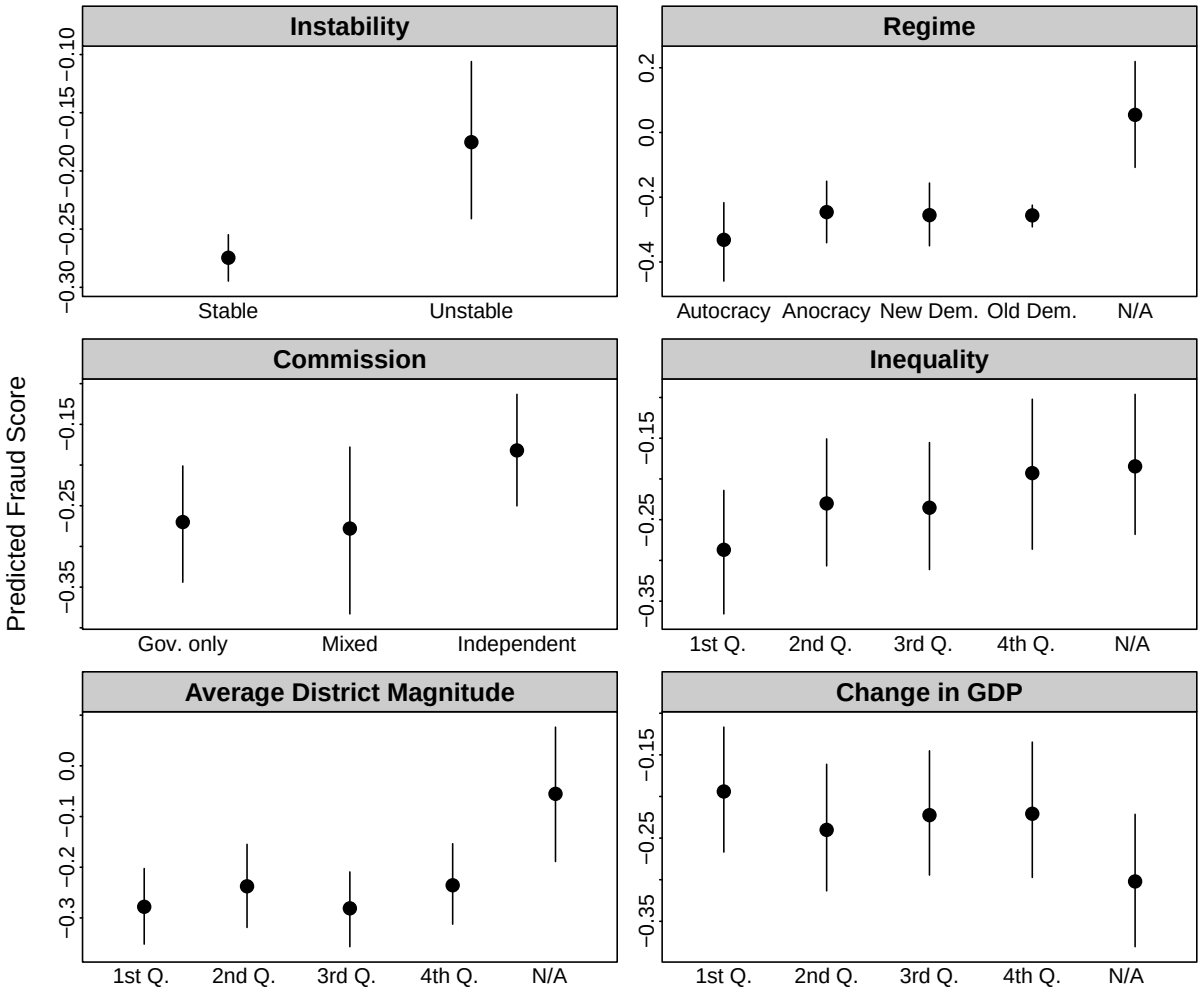
The points indicate the average inclusion probabilities for each variable in a BART model fit with ten trees ($n = 598$). Higher values indicate the variable was used more often in creating splitting rules in the binary trees (see Footnote 15). Economic inequality, average district magnitude, regime type, change in GDP, and age of democracy have been transformed into categorical variables to accommodate missing values.

Figure 3. Partial dependence plots for seven continuous variables



Each panel shows the expected value (solid line) and 80% credible intervals (shaded area) for the expected value of the fraud proxy as the corresponding covariate spans its observed range. These quantities are estimated for each observation across the entire posterior while holding all other covariates at their observed value (Chipman, George and McCulloch, 2010).

Figure 4. Partial dependence plots for six categorical variables



Each panel shows the expected value (solid point) and 80% credible intervals (vertical bars) for the expected value of the fraud proxy at each observed value of the corresponding covariate. These quantities are estimated for each observation across the entire posterior while holding all other covariates at their observed value (Chipman, George and McCulloch, 2010).

our proxy, as unstable polities are estimated to have significantly higher scores. Similarly, electoral fraud is more common in countries with extremely high levels of ethnolinguistic fractionalization, as evidenced by the top left panel of Figure 3.

Surprisingly, while urbanization appears to be a relevant predictor of fraud, the top right panel of Figure 3 shows that urbanization appears to have a largely *negative* relationship with our fraud proxy. Low levels of urban population density are most associated with high scores on our fraud proxy, while more urban countries are predicted to have lower values. Although this runs counter to what previous studies have found (e.g. Domínguez and McCann, 1998), it is consistent with the idea that local political machines are more successful at engaging in fraudulent activities in rural settings than in urban and settings, where electoral observers can more easily pick up on such strategies (Lehoucq, 2003).[17]

Our model also uncovers a non-linear relation between turnout and fraud, as very low rates of voter turnout are predicted to have high levels of fraud as captured by our NELDA-based proxy, average rates (50% to 90%) are not indicative of it, and very high rates of turnout (above 90%) are associated with relatively high levels of our proxy. The relationship largely coheres to our theoretical expectation, with the exception of the effect of very high turnout. However, beyond the 90% mark, there is a simple explanation: even in countries with compulsory voting laws, turnout in excess of 90% is rare (Blais, 2006), suggesting that such results should raise a red flag.

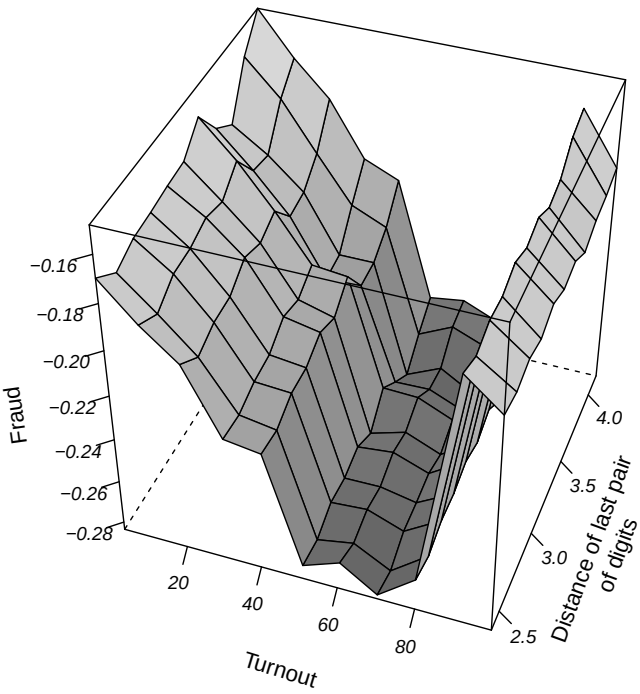
Forensic indicators: In contrast, Figure 3 shows that – considered independently – the forensic indicators have effects that are substantively modest and confidence intervals that are sufficiently wide to suggest they their predicted relationship with fraud independent of all other measures is difficult to distinguish from a null effect.[18] However, this is not a surprising result. In line with our argument that forensic indicators should be adjusted to account for context, the role of the forensic indicators of anomalous digit distributions in the BART model is better understood by examining forensic and contextual factors interactively.

To illustrate this point, we calculated two-way partial dependence plots for the most important forensic indicator (the average distance between the last two digits), and one of the most important contextual risk-factors (turnout). Figure 5 shows that the the degree to which a distance between the last pair of digits is indicative of fraud depends on the level of electoral turnout: while a very high distance are associated with higher values of fraud, the forensic tool loses almost all of its discriminating ability at the mid-level ranges. For

[17] Also surprisingly, independent electoral commissions are associated with higher values of our fraud proxy – though this may be the result of a reversed relationship: such commissions are likely to be installed in places where electoral fraud is prevalent.

[18] The wider 80% credible intervals at each end reflect the lower number of observations at these extremes.

Figure 5. Interactive partial dependence for forensic tools and contextual risk factors



Interaction between turnout level and the average distance between the last pair of digits in district-party vote totals. The forensic indicator is most informative about the fraud proxy for low levels of turnout.

instance, when turnout takes on its minimum observed value, a change from 3.8 to 4.1 on the forensic indicator is associated with an increase of roughly 0.03 in our fraud proxy. When turnout held at 60%, however, the same change in the forensic indicator is associated with a change of 0.01 in the fraud proxy. While these effects are nowhere near in magnitude to those of contextual indicators, they are important insofar as they illustrate the benefits of using forensic tools alongside contextual risk factors.

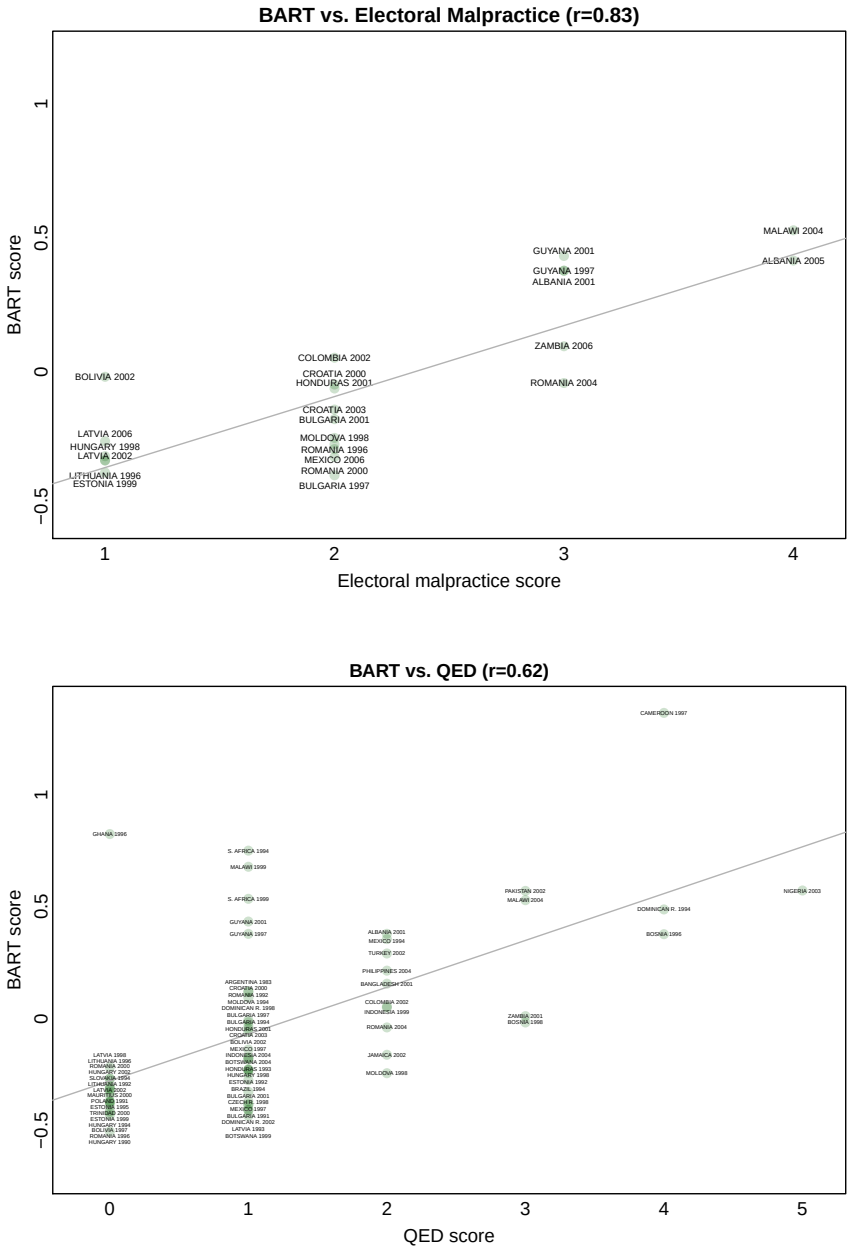
4.4. *Assessing validity*

To further validate our fraud proxy and informed forensics approach for identifying it, we compared the predictions from our full BART model to two other notable sources for cross-national data on fraud: the *Database on Electoral Malpractice* (Birch, 2007, 2012) and the *Quality of Elections Database (QED)* (Kelley, 2012). These projects take quite different approaches to identifying instances of electoral fraud.[19] If the predictions from our combined BART model correlate with these measures, it will give us further confidence that the informed forensics approach is a valid method for identifying fraud. For each measure, we focus only on elections where available BART estimates overlap with observations by the alternative measures. This leaves us with 23 and 64 observations for the *Electoral Malpractice* and *QED* measures of fraud, respectively.

As shown in Figure 6, the predictions from the model correlate well with both alternatives, although the correspondence is obviously imperfect. The top panel of Figure 6 plots predictions from the BART model against the *Electoral Malpractice* score. The correlation here is quite strong ($r = 0.83$), although our BART prediction tends to group a higher proportion of the overall observations in the least-likely fraudulent category. The bottom panel plots the BART predictions against the QED measure. This correlation is more modest (0.62), but far from trivial. Here again, relative to this alternative measure, our model's predictions tend to place most cases in the "least fraudulent" category. Nevertheless, we interpret these results as supportive of our informed forensics approach generally, although reflecting some of the inherent limitations of available data for training the model.

[19] That is, both databases hand code the results of election observer mission reports across a variety of on-the-ground observable outcomes, i.e. ballot stuffing, problems with election administration, voter intimidation, and so forth. Based on these eye-witness accounts, Birch's database codes the overall quality of the election on a 4-point scale. Kelley's database codes two related 3-point assessments, which we combine linearly to provide an overall portrait of the likely incidence of fraud: the *degree* of electoral misconduct and the *extent* of electoral misconduct. In both cases, higher scores indicate increasingly problematic elections and, thus, both metrics should be positively correlated with the predictions from our BART model. More information on these two metrics can be found in Appendix C.

Figure 6. Comparing predictions from the BART model with alternative measures



Each panel show the correspondence of the model predictions with measures from the *Database on Electoral Malpractice* ($n = 23$) and the *Quality of Elections Database* ($n = 64$) (Birch, 2012; Kelley, 2012). The points show the position of the country-year observations and the lines show the bivariate linear regression relationship.

5 CONCLUSION

Electoral fraud is notoriously difficult to detect. Election monitoring puts "boots on the ground," but resource limitations and selection biases prevent researchers from gaining a clear picture of the nature and extent of fraud as identified by monitors. Context-rich case studies of elections suffer from problems of scope and generalizability. For example, the existence of regional strong men may lead to specific predictions about fraud in *this* country (say, Russia), but have little to say about fraud in *some other* country (say, Argentina). Finally, the methodology that is most readily implementable cross-nationally – electoral forensics – has been developed various mathematical indicators, all of which have something to tell researchers, but none of which is a "silver bullet" in isolation.

The informed forensics approach we propose is designed to build on the strengths of extant research on electoral fraud by ameliorating some the most obvious weaknesses of the forensics literature and systematically incorporating insights from the rich substantive literature on socio-demographic and institutional factors that create contexts where fraud is more likely. First, the procedure provides a seamless – and systematic – method for adjusting purely forensic indicators to specific electoral contexts. The informed forensic approach outlined above amalgamates forensic and context-rich indicators into a single statistical model in a flexible manner. Specifically, the BART model allows for high-level interactions such that, for example, unusual digit distributions are deemed a more credible indicator of fraud in countries with authoritarian tendencies. Further, the BART model allows for the possibility that relationships may be nonlinear so that, for instance, the χ^2_{2BL} statistic (i.e. the test statistic measuring deviations from the theoretical distribution of the second digit values according to Benford's Law) may only become indicative of fraud at extremely high values.

Second, by building the model on a large cross-national database, the degree to which any specific forensic indicator, contextual risk factor, or combination thereof is weighted in the model is based on its empirical performance. The researcher's primary role is in selecting a set of features – forensic or contextual – to include in the analysis. BART then offers sufficient statistical flexibility that it (largely) lets the data speak for itself insofar as we – as researchers – are not bringing to the modeling process our *a priori* expectations about which forensic indicators or risk factors are the most salient determinants of fraud. Accordingly, the BART model utilizes electoral returns and contextual factors to identify fraud only insomuch as specific patterns (e.g., digit distributions or digit distributions in combination with contextual characteristics) are valid empirical indicators of fraud cross-nationally. Because our BART model has been fit on the distribution of election characteristics across several hundred elections, each new out-of-sample election can be evaluated against this broader international backdrop. The truly comparative nature of the fraud assessment here is a marked

improvement over methods that attempt to evaluate a single country's election outcome in isolation.

While our model was built on a large cross-national database, there is no reason why it could not be used within a single country to evaluate whether electoral returns vary in their abnormality across, say, American states or Ukrainian *oblasts*. As long as scholars or practitioners can obtain observations across multiple units at the sub-geographic unit level, they can be fed into a tree-based model just as we have fed in multiple districts at the sub-national level. The only additional challenge such an approach would have is that any context-rich features one wished to include in the trees' approach would have to be observed at that state or *oblast* level. In other words, we could compare returns in precincts within one *oblast* to returns in precincts another *oblast* as long as we could observe items like, for example, change in GDP per capita at the *oblast* level.

Empirically, our expectations about the benefits of combining contextual and forensic indicators are supported by our analyses. First, we demonstrated that the combined out-of-sample predictions from BART models fit using *both* forensic and contextual features provide better out-of-sample predictive power than either method does individually. Second, we showed that BART does, in fact, rely on a mix of forensic and contextual factors in generating its predictions and that these factors do interact in a highly non-linear fashion. Finally, we showed that the estimates from the BART model mirror (if imperfectly) estimates of electoral fraud generated from alternative methodologies. Indeed, our informed forensic indicators correlate modestly with these alternative measures, despite the fact that our indicators are far less resource intensive in implementation.

Given the ability of this informed forensics approach to identify variations in the degree of voting irregularities with relatively easy to obtain data, we believe it has the potential to represent a parsimonious but effective "off-the-shelf" method of identifying potential voting irregularities cross-nationally – even in cases where outside observers were not present at the election. Nonetheless, it is important to note that the tool as presented above is certainly not itself a "silver bullet" for infallibly identifying fraud. First and foremost, our results are necessarily constrained by available data, and particularly the lack of election data reported at lower levels of aggregation (e.g., precincts) cross-nationally. It seems reasonable to conjecture that some of the forensic indicators above may perform better if more fine-grained election-level data was available for more countries. Further, the set of countries that provide even district-level data is significantly censored, meaning that further research is needed to validate this to a broader set of electoral circumstances as additional data becomes available.

Second, it may not be possible for any statistical method to provide a definitive statement

of maliciously-motivated fraud – after all, irregular patterns in vote returns could occasionally be the result of other types of (non-malicious) tabulation problems. Third, although the BART model does have several specific advantages, there are obviously several alternative machine learning techniques in the literature which may result in somewhat different findings. Finally, the advantage of semi-supervised machine learning algorithms such as BART is that they are able to uncover complex relationships between sets of covariates and outcomes of interest. The disadvantage, however, is that the results depend on the degree to which the outcome of interest has been measured appropriately. Future research, therefore, may seek to develop alternative proxy measures of fraud to incorporate into the training of an informed forensics framework.

Despite these limitations, we argue that this general approach may be of use to political practitioners, journalists, and aid agencies due to its easily implementable nature. While we may not be able to declare an election fraudulent based solely on the model above, the approach can certainly identify elections whose results are worthy of greater scrutiny and suspicion as a result of contextual risk factors, unusual vote returns, or some combination of thereof. Further, due to the fact that our BART model has been fit based on the distribution of election characteristics *across several hundred elections*, each new out-of-sample election can be evaluated against this broader international backdrop. Finally, as more and better data become available, we anticipate that the utility and generalizability of the informed forensic approach will improve.

5 REFERENCES

Acemoglu, Daron and James A. Robinson. 2006. *Economic Origins of Dictatorship and Democracy*. Cambridge Univ Press.

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

Beck, Thorsten, George Clark, Alberto Groff, Philip Keefer and Patrick Walsh. 2001. "New Tools in Comparative Political Economy: The Database of Political Institutions." *World Bank Economic Review* 15(1):165–176.

Benford, Frank. 1938. "The Law of Anomalous Numbers." *Proceedings of the American Philosophical Society* 78(4):551–572.

Birch, Sarah. 2007. "Electoral Systems and Electoral Misconduct." *Comparative Political Studies* 40(12):1533–1556.

Birch, Sarah. 2012. *Electoral Malpractice*. Oxford, UK: Oxford University Press.

Blais, André. 2006. "What Affects Voter Turnout?" *Annual Review of Political Science* 9:111–125.

Boix, Carles. 1999. "Setting the Rules of the Game: The Choice of Electoral Systems in Advanced Democracies." *American Political Science Review* 93(3):609–624.

Brancati, Dawn. 2007. *Constituency-Level Elections (CLE) Dataset*. New York: Constituency-Level Elections Dataset. URL: http://www.cle.wustl.edu.

Brandt, Patrick T., John R. Freeman and Philip A. Schrodt. 2014. "Evaluating Forecasts of Political Conflict Dynamics." *International Journal of Forecasting* 30:944–962.

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

Chipman, H.A., E.I. George and R.E. McCulloch. 2010. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics* 4(1):266–298.

Cho, W.K. Tam and B.J. Gaines. 2007. "Breaking the (Benford) Law." *The American Statistician* 61(3):218–223.

Cox, Gary W. 1999. "Electoral Rules and the Calculus of Mobilization." *Legislative Studies Quarterly* 24:387–419.

Cox, Gary W. and J. Morgan Kousser. 1981. "Turnout and Rural Corruption: New York as a Test Case." *American Journal of Political Science* 25(4):646–663.

Dardé, Carlos. 1996. Fraud and Passivity of the Electorate in Spain, 1875-1923. In *Elections Before Democracy: The History of Elections in Europe and Latin America*, ed. Eduardo Posada-Carbó. MacMillan Press pp. 201–223.

Domínguez, Jorge I. and James A. McCann. 1996. *Democratizing Mexico: Public Opinion and Electoral Choices*. Baltimore, MD: Johns Hopkins University Press.

Domínguez, Jorge I and James A McCann. 1998. *Democratizing Mexico: Public opinion and electoral choices*. Baltimore: Johns Hopkins University Press.

Efron, Bradley and Robert Tibshirani. 1997. "Improvements on Cross-Validation: The 632+ Bootstrap Method." *Journal of the American Statistical Association* 92(438):548–560.

Friedman, Jerome, Trevor Hastie and Rob Tibshirani. 2010. "Regularization paths for generalized linear models via coordinate descent." *Journal of statistical software* 33(1):1.

Green, Donald P and Holger L Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." *Public opinion quarterly* 76(3):491–511.

Grendar, Marian, George Judge and Laura Schechter. 2007. "An Empirical Non-Parametric Likelihood Family of Data-Based Benford-Like Distributions." *Physica A* 380:429–438.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hill, Jennifer. 2012. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 10(2):217–240.

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

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

Kitschelt, Herbert, Zdenka Mansfeldova, Radoslaw Markowski and Gábor Tóka. 1999. *Post-Communist Party Systems: Competition, Representation, and Inter-Party Cooperation*. Cambridge, UK: Cambridge University Press.

Kollman, Ken, Allen Hicken, Daniele Caramani and David Backer. 2011. *Constituency-Level Elections Archive (CLEA)*. Ann Arbor, MI: University of Michigan Center for Political Studies.
URL: *www.electiondataarchive.org*

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

Levitsky, Steven and Lucan A. Way. 2002. "Elections Without Democracy: The Rise of Competitive Authoritarianism." *Journal of Democracy* 13(2):51–65.

Martín, Isbella. 2011. "2004 Venezuelan Presidential Recall Referendum (2004 PRR): A Statistical Analysis from the Point of View of Electronic Voting Data Transmissions." *Statistical Science* 26(4):528–542.

Mebane, Walter R. 2008. Election Forensics: The Second-Digit Benford's Law Test and Recent American Presidential Elections. In *Election Fraud: Detecting and Deterring Electoral Manipulation*, ed. R. Michael Alvarez, Thad E. Hall and Susan D. Hyde. Washington, D.C.: Brookings Institute Press.

Mebane, Walter R. 2010. "Fraud in the 2009 Presidential Election in Iran?" *Chance* 23(1):6–15.

Mebane, Walter R. 2012. "Second-digit Tests for Voters' Election Strategies and Election Fraud." Paper presented at the 2012 Annual Meeting of the Midwest Political Science Association, Chicago.

Mebane, Walter R. and Kirill Kalinin. 2009. "Comparative Election Fraud Detection." Prepared for the Annual Meeting of the American Political Science Association.

Schedler, Andreas. 2002. "The Menu of Manipulation." *Journal of Democracy* 13(2):36–50.