Political Analysis (2011) 19:409–433
doi:10.1093/pan/mpr033

Fraudulent Democracy? An Analysis of Argentina's *Infamous Decade* Using Supervised Machine Learning

Francisco Cantú and Sebastián M. Saiegh

Department of Political Science, University of California, San Diego, CA 92093
e-mail: ssaiegh@ucsd.edu (corresponding author)

In this paper, we introduce an innovative method to diagnose electoral fraud using vote counts. Specifically, we use synthetic data to develop and train a fraud detection prototype. We employ a naive Bayes classifier as our learning algorithm and rely on digital analysis to identify the features that are most informative about class distinctions. To evaluate the detection capability of the classifier, we use authentic data drawn from a novel data set of district-level vote counts in the province of Buenos Aires (Argentina) between 1931 and 1941, a period with a checkered history of fraud. Our results corroborate the validity of our approach: The elections considered to be irregular (legitimate) by most historical accounts are unambiguously classified as fraudulent (clean) by the learner. More generally, our findings demonstrate the feasibility of generating and using synthetic data for training and testing an electoral fraud detection system.

1 Introduction

How can we distinguish an electoral landslide from a stolen election? As the controversy over the 2009 Iranian presidential contest illustrates, it is often difficult to determine whether an election is fraudulent. Mr. Ahmadinejad's wide margin of victory coupled with the speed with which the results were certified led many to doubt the authenticity of the outcome. Indeed, within days of the election, a series of reports analyzed a number of statistical anomalies and concluded that the election was rigged (Roukema 2009; Zetter 2009). Most experts, however, agreed that while there was little reason to trust the election results, the evidence furnished by these studies was inconclusive (Mebane 2010; Zetter 2009).[1]

Establishing whether an electoral outcome is the reflection of voters' preferences or the result of manipulation is of utmost importance: Elections are essential mechanisms for providing public accountability, transparency, and representation. As Alvarez, Hall, and Hyde (2008) note, fraud and manipulation corrupt this process and prevents voters' voices from being heard. But, as they also point out, despite its importance, little is known about electoral fraud.[2]

The operational problem in uncovering fraudulent elections is identifying the characteristics that make them distinct from valid ones. It is usually impossible, however, to be absolutely certain about the legitimacy of an election. A particularly appealing option to overcome this difficulty is to tease out possible evidences of fraud from the available data using mathematical algorithms. For example, recent studies examine the distribution of the digits in reported vote counts to detect fraudulent practices (Pericchi and Torres 2004; Mebane 2006, 2008b; Beber and Scacco 2008; Roukema 2009).[3]

[1] For example, Nate Silver wrote in his FiveThirtyEight blog, "... I have no particular reason to believe the results reported by the Interior Ministry. But I also do not have any particular reason to disbelieve them, at least based on the statistical evidence ..." (posted on June 13, 2009). See also Andrew Gelman's post of June 17, 2009 in the same blog.

[2] See Lehoucq (2003) for an extensive review of the literature on electoral fraud.

[3] Digit analysis using Benford's Law is an example of such a method. Benford's Law specifies that in a collection of numbers, the first possible digits should not occur with equal frequency. Widely applied to financial auditing (Drake and Nigrini 2000), conformity with Benford's Law has also been used to detect manipulation of economic indicators (Nye and Moul 2007), campaign

This approach, however, is not without its drawbacks. Digit tests are essentially based on comparing observed data with expected values, but expected values can be derived in various ways, depending on the context.[4] In addition, electoral manipulation can be perpetrated in many different ways. Hence, in order to assess the benefits of digital analysis as a useful fraud-detection tool, numerous data sets with known levels of deviation from a given digit distribution would be needed (Busta and Weinberg 1998). From a practical standpoint, things are further complicated by the fact that it may be impossible or at least very difficult to acquire the amount or type of data needed for such tests: *Authentic* data sets of vote counts with fraudulent entries are rarely available, and we usually have no control over what type of fraudulent practices the data may contain.[5]

In this paper, we introduce a novel method to diagnose electoral irregularities using vote counts. Specifically, we propose the use of machine learning techniques to detect fraud. A typical supervised learning problem is comprised of two components: (1) an outcome measurement, which can be either quantitative or categorical (such as fraudulent/not fraudulent elections) and (2) a *training set* of data, which includes the outcome and *feature* measurements for a set of objects (such as electoral contests). Given these data, two standard tasks are to build a *learner* that most accurately predicts the class of a new example (classifier design) and to identify a subset of the features that is most informative about the class distinction (feature selection) (Hastie, Tibshirani, and Friedman 2009).[6]

Taking into account the data availability problems mentioned above, our main innovation is the use of synthetic data to develop and train an electoral fraud detection prototype. First, we use Monte Carlo methods to generate large amounts of electoral data that preserve the statistical properties of a selected set of authentic data used as a seed. Next, we employ a naive Bayes classifier as our learning algorithm and rely on digital analysis to identify the features that are most informative about class distinctions. The specific procedure is the following: (1) we create a set of simulated elections. This *training* set is composed of two disjoint subsets: one containing vote counts that follow a distribution with known properties and another where the data are purposively "manipulated"; (2) we rely on digital analysis for feature selection and then calibrate membership values of the simulated elections (i.e., clean/manipulated) using logistic regression; (3) we recover class-conditional densities using the relative frequencies from the training set; (4) we evaluate the detection capability of the classifier using authentic data drawn from a novel data set of district-level vote counts in the province of Buenos Aires (Argentina) between 1931 and 1941, a period with a checkered history of fraud. Our results corroborate the validity of our approach: The elections considered to be irregular (legitimate) by most historical accounts are unambiguously classified as fraudulent (clean) by the *learner*.

The paper makes two important contributions. One is methodological. Our findings demonstrate the feasibility of using synthetic data for training and testing an electoral fraud detection system. Building a learner using simulated rather than real data is quite common in automated fraud detection research (Phua et al. 2005). For example, Wong et al. (2003) use Bayesian networks to uncover simulated anthrax attacks from real emergency data.[7] Synthetic data are less frequently used in financial fraud detection; still, the use of manipulated data is discussed in some papers (Busta and Weinberg 1998; Chan et al. 1999). To our best knowledge, though, this paper represents the first study in political science that uses synthetic data to develop and train a fraud detection prototype for electoral contests.

finances (Cho and Gaines 2007), and survey data (Schäfer et al. 2004). Studies using Benford's Law to analyze voting counts include the analysis of elections in Venezuela (Pericchi and Torres 2004), Russia (Mebane 2008a), Iran (Mebane 2010; Roukema 2009), Mexico (Mebane 2006), and the United States in both contemporary and Gilded Age elections (Mebane 2006, 2008b; Buttorf 2008).

[4] For example, election returns are unlikely to follow Benford's Law where districts are of nearly equal size and the level of competition fixes most vote totals in a limited range (Cho and Gaines 2007).

[5] For a comprehensive discussion of the drawbacks associated with the study of electoral fraud using observational rather experimental data, see Hyde (2007).

[6] Learning is called supervised, in contrast to *unsupervised*, because the learning process is guided by an outcome variable. In the unsupervised learning problem, only features are observed and no outcome measurements exist (Hastie, Tibshirani, and Friedman 2009).

[7] Within computer science, significant work has been done using synthetic test data in network intrusion detection research (Puketza et al. 1996; Debar et al. 1998; Kvarnstrom, Lundin, and Jonsson 2000; Haines et al. 2001). Likewise, the use of simulated data to train a *learner* is also a standard practice in bioinformatics and molecular biology (Demichelis et al. 2006), as well as in antiterrrorism, law enforcement, and other security areas.

Our second contribution is substantive. After training the *learner* with the synthetic data, we used it to detect the fraudulent contests employing a novel data set of district-level vote counts in the province of Buenos Aires between 1931 and 1941. The conventional wisdom states that electoral fraud, rather than a change in voter preferences, led to the dramatic electoral shifts during this period—which is known in Argentine politics as the "infamous decade" (Alston and Gallo 2010). Most studies, however, use anecdotal evidence, journalistic accounts, or fraud complaints. The comparison of the detection results with authentic data confirms the validity of the conventional wisdom. An important virtue of our approach, though, is that such validation is based on the distribution of the digits in reported vote counts rather than on the perceptions of certain actors. As Alston and Gallo (2010) note, the infamous decade produced a watershed in Argentine politics leading to structural political and economic changes in its aftermath. As such, demonstrating by means of our proposed method, the presence of fraud during this period constitutes another very important contribution to the literature.

The remainder of this paper is organized as follows. In Section 2, we describe the authentic and synthetic data. Section 3 introduces the naive Bayes classifier. In Section 4, we present our main empirical findings. A final section concludes.

2 Electoral Fraud: Authentic and Synthetic Data

In order to train a *learner*, a set of input and output objects, known as a training set, should be gathered. Specifically, for any given problem, we need a set of variables, denoted as *inputs*, or *features* (i.e., the *independent variables*), which are measured or preset. These should have some influence on one or more *outputs* (i.e., the *responses* or the *dependent variables*).

Given these data requirements, one needs to take into account two major considerations when building an electoral fraud detection system. The first is the problem of unbalanced class sizes: Data on legitimate elections usually outnumber information regarding fraudulent ones. The second concerns the uncertainty of class membership (fraudulent elections may remain unobserved and thus be labeled legitimate). These two problems are obviously interrelated: The extent of electoral fraud is difficult to quantify mostly because governments seldom advertise the fact that they have cheated (Przeworski 2010).

One alternative in overcoming these problems is to create synthetic data (Clifford and Heath 1993; Rubin 1993; Katz and Sala 1996; Lundin, Kvarnström, and Jonsson 2002; Reiter 2004; Eno and Thompson 2008). Using synthetic data for evaluation, training, and testing, a supervised learning model offers several advantages over using authentic data. First, we can generate more data sets than what would be available using only real data. Second, properties of synthetic data can be tailored to meet various conditions which may not be clearly observable in the real data. Third, variations of known frauds (or new frauds) can be artificially created to study how these changes affect performance parameters, such as the detection rate (Lundin, Kvarnström, and Jonsson 2002).

Synthetic data, though, should be realistic; that is, they should reflect selected properties of the real data. Thus, as a first step, one should collect authentic data and identify the key parameters that must be preserved in the synthetic samples. The next section describes how the authentic data were collected. We then discuss the synthetic data generation method and their conformity to the actual data in Sections 2.2 and 2.3, respectively.

2.1 *Authentic Data: Elections in Buenos Aires (1931–1941)*

Mass electoral participation in Argentina began with the passage of the Sáenz Peña Law in 1912, which established the secret ballot and mandatory voting. These reforms ended a long period of oligarchic republicanism. In the following decades, two main parties alternated power at the national level. These were the Partido Conservador, or Conservative party, and the Unión Civica Radical (UCR or Radical party). Conservatives controlled the government until the election of Radical Hipólito Yrigoyen in 1916.

The Radical years ended in 1930 when a Conservative-backed military coup ousted Yrigoyen from office following his reelection to the presidency in 1928. The coup was followed by an election in 1931 that restored power to a conservative coalition. During the remainder of the 1930s, the Conservatives continually used electoral fraud to maintain power. As José Luis Romero (1959) notes, fraud and privilege were the main characteristics of this period, which became known in Argentine politics as the "infamous decade" (Ciria 1974).

Radicals and Conservatives also vied for electoral allegiance from 1931 to 1942 in Buenos Aires. Covering an area of 188,446 square miles, Buenos Aires is Argentina's largest province. A third contender, the Partido Socialista (PS or Socialist party), was well organized and enjoyed some strength in suburban and coastal districts but never gained a firm foothold in the countryside (Walter 1985). In addition, elections during this period were held under the *lista incompleta*, or incomplete list electoral system, which discriminated against third-party representation: Ballots included a list of candidates for at least two-thirds of the positions at stake. Two-thirds of the seats were assigned to the winning list, the remaining third went to the list that followed in number of votes (Abal Medina and Cao 2003). In consequence, the Socialists seldom managed to gather more than 10% of the vote.[8]

Electoral contests were held regularly during this period; however, according to most historical accounts, some of them were far from regular (Drake 2009). Accusations of fraud and corruption were constant characteristics of many elections, and few contests passed without complaints from one side or the other. Anecdotal evidence suggests that electoral manipulation varied greatly: Ballot boxes were stuffed; the dead rose to vote on election day; police and local officials often intimidated, harassed, and coerced potential opposition voters; polls were opened late and closed early; and government employees and others traveled the province to vote numerous times in the same election (Walter 1985; Bejar 2005).

Although all of these practices can affect election outcomes, an important distinction needs to be made between fraud and manipulation. For example, the use of repeaters, the election-day importation of voters from other provinces, "vote buying" (including both paying individuals for switching their vote choices as well as rewarding opponents for not voting), or more legitimate activities, such as providing free transportation to the polls or organizing a massive get out the vote campaign, can alter vote totals in an election (Cox and Kousser 1981).

Yet, to be absolutely clear, what characterized electoral fraud in the province of Buenos Aires during this period were the efforts by the Conservatives to systematically disenfranchise voters. Its "gangster toughs," in the words of the *Review of the River Plate*, would ward off the Radical voters from the polling stations by brute force and intimidation. These "missing" votes were then conveniently replaced by Conservative ballots.[9]

The opposition's denunciation of electoral fraud in the province of Buenos Aires at the outset of the "infamous decade" provides a good summary of these practices:

> "In addition to those citizens who were compelled to vote in a certain way, those who were stripped of their identification cards, and those who made it to the polls only to suffer harassment or punishment, those individuals who decided to stay home to avoid violence also ended up 'voting'. . . ."[10]

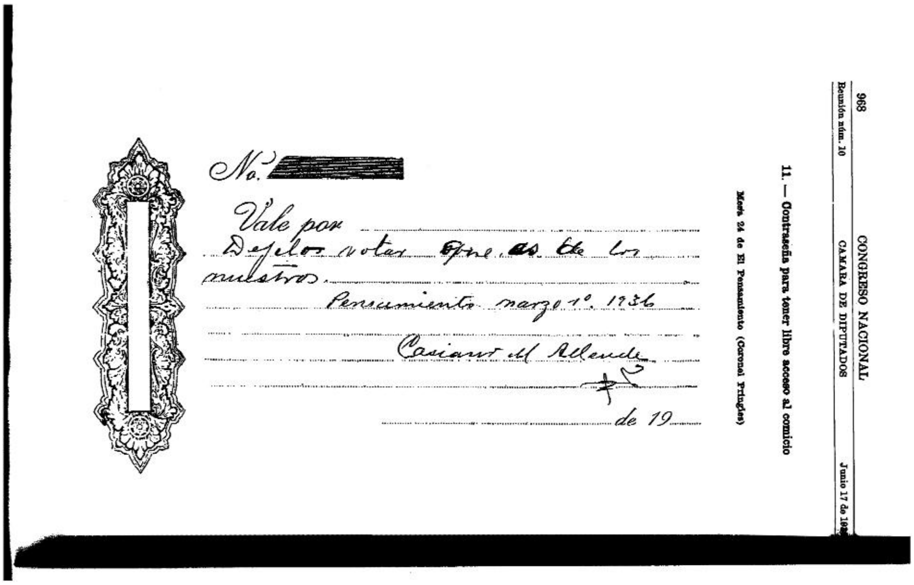
The document reproduced in Fig. 1 provides a good illustration of what constituted fraud in this setting. Introduced in the official records of the Argentine Chamber of Deputies as evidence of electoral tampering, the document shows how written "passes" were used to give admission to the polls to some voters but not others. The handwritten directions in the permit say ". . . Let him vote, he is one of us . . . ," giving the bearer authorization to access polling station number 24 located in El Pensamiento (Coronel Pringles).

The document also exemplifies two additional features of electoral malfeasance in the 1930s. First, most allegations of fraud and coercion occurred during the periods of Conservative control of the province, and they usually involved accusations that votes were taken away from the Radical opposition and given to the Conservatives. Second, irregularities seemed to occur in almost every polling station throughout the province of Buenos Aires. In other words, it appears that the falsification of the popular will was not restricted to a few "decisive" wards and/or polling stations but instead was a widespread phenomenon (Ciria 1974; Walter 1985; Bejar 2005).

[8] The Socialists obtained 9.0% of the vote in the April 5, 1931 election; 3.6% in the November 3, 1935 election; 5.7% in the election held on March 1, 1936; 3.3% in the March 3, 1940 election; and 2.7% in the election held on December 7, 1941. The only two contests in which the Socialists obtained more than 10% of the vote were the ones in which the Radical party did not participate (the November 8, 1931 and the March 4, 1934 elections).

[9] "Awaiting the Verdict," in *The Review of the River Plate*, March 20, 1936, pages 6–7.

[10] Cámara de Diputados. *Diario de Sesiones*, July 20–21, 1932, III: 232.



968 CONGRESO NACIONAL Junio 17 de 1936

Reunión núm. 10 CAMARA DE DIPUTADOS

11. — Contraseña para tener libre acceso al comicio

Mesa 24 de El Pensamiento (Coronel Pringles)

No. [redacted]

Vale por

Dejelos votar que es de los nuestros.

Pensamiento marzo 1°. 1936

Casiano M Allende

de 19

Fig. 1 Written authorization to access a polling station. This document provides a good illustration of the fraudulent practices that took place in the province of Buenos Aires during the "infamous decade": written "passes" were often used to give admission to the polls to some voters but not others. The handwritten directions in the permit say "… Let him vote, he is one of us …," giving the bearer authorization to access polling station number 24 located in El Pensamiento (Coronel Pringles). Source: Argentine Chamber of Deputies, *Diario de Sesiones*, June 17, 1936: 968.

Evidence indicating that other forms of electoral finagling (such as manipulating turnout) mattered is not persuasive. For example, no significant differences in turnout rates for the 1936, 1940, and 1941 elections exist (they amounted to 61.0%, 60.4%, and 59.2%, respectively). Yet, these elections had diametrically different outcomes (see Section 2.1.1 below). The analysis of the electoral returns also suggests that violence and intimidation, rather than strategic voting, characterized Buenos Aires' elections in this era. As noted above, the electoral rules strongly encouraged a *Duvergerian* competition: According to the so-called "wasted-vote logic," a voter who would otherwise favor a third-party candidate can increase her utility by voting for one of the parties most likely to be tied for first (Cox 1997). It seems highly unlikely that Socialists voters would strategically abandon their party in favor of Conservative candidates. Hence, we should expect fewer votes to be "wasted" when the Conservative party is the runner up. The Socialist vote, however, did not exhibit considerable fluctuations in the electoral contests under consideration.[11]

In sum, electoral outcomes can be manipulated in a number of ways. The historical evidence, though, suggests that electoral fraud in the province of Buenos Aires during this period consisted mostly of switching a proportion of votes from one party to another.

2.1.1 Data description

The conventional wisdom states that electoral fraud led to the notable electoral shifts that occurred in the province of Buenos Aires during the "infamous decade." Yet, the scope, intensity, and efficacy of fraud are not well documented in the literature. As discussed above, a particular challenge in electoral fraud research is data availability. Indeed, most of the existing accounts use anecdotal evidence, press reports, or fraud complaints. Nonetheless, no study thus far has used microlevel electoral data for the period between 1931 and 1941.[12]

[11] See footnote 8 above.

[12] Lupu and Stokes (2009) assembled an impressive database of Argentine elections between 1912 and 2003. Their sample, however, does not include most of the elections that took place during the 1931–1941 period, and the geographical coverage of the ones that are included in their sample (the elections of 1937 and 1940) is quite small and excludes the province of Buenos Aires.

We collected a complete record of the gubernatorial elections of 1931, 1935, and 1941, as well as the 1936 and 1940 national congressional elections in the province of Buenos Aires.[13] The data were obtained from two official sources of information: the *Memorias* of the Ministerio del Interior and the *Diario de Sesiones* of the Argentine Chamber of Deputies. The units of our analysis are the electoral returns in each of the Buenos Aires' *partidos*. The partido is the basic local administrative unit of the province. As Walter (1985) notes, most of these partidos bear the same name as the largest municipality within their confines. These municipalities are somewhat analogous to county seats in the United States.

Between 1931 and 1941, some partidos experienced name changes and shifting boundaries, but the total number of districts remained at 110. We exclude the following ten partidos from the analysis: General Alvear, General Conesa, General Lavalle, and Pila as well as Avellaneda, Bahia Blanca, La Plata, Lomas de Zamora, Quilmes, and San Martin. The former group consists of partidos where at least one of the parties received an insignificant number of votes (e.g., the Radicals received only two votes in Pila in the 1935 elections and the Conservatives only one vote in General Alvear in 1940). In contrast, the latter group is comprised of densely populated urban centers. For example, in the provincial capital of La Plata, a bureaucratic-university city, a total of 32,929 votes were cast for the two main parties in 1931 and 43,426 ten years later. According to Walter (1985), given their large number of registered voters and their degree of urbanization, electoral fraud was quite rare in these partidos. We thus observe district-level vote counts for each of the five aforementioned elections in 100 partidos. Table 1 provides an overview of the electoral data.

The entries in Table 1 show the distribution of votes for the Radical (UCR) and Conservative parties. These figures underscore the reversal of electoral outcomes. In the 1931 elections, the average vote for the Radicals at the partido level was 1673 versus 1384 for the Conservatives. In 1935, they lost to the Conservatives by a margin of 2-to-1 (2099 votes versus 1057). As Walter (1985) notes, some of the numerical reverses at the partido level were ludicrous. In 1931, for example, the Radicals had won General Sarmiento by a vote of 1573 to 1341; in 1935, the Conservatives carried the partido by a count of 3000 to 40. A comparison between the 1940 and 1941 elections reveal a similar pattern.

Table 1 Elections in the province of Buenos Aires (1931–1941)

	Observed	*Mean*	*Standard deviation*	*Minimum*	*Maximum*
1931					
UCR	100	1672.81	1007.76	262	4636
Conservatives	100	1384.30	778.01	339	4081
1935					
UCR	100	1057.08	957.458	40	4641
Conservatives	100	2099.71	1332.41	554	6772
1936					
UCR	100	1212.42	965.89	81	5060
Conservatives	100	1689.55	1094.59	395	5576
1940					
UCR	100	1874.02	1246.69	264	6103
Conservatives	100	1470.81	896.30	411	3998
1941					
UCR	100	1292.91	911.17	224	4979
Conservatives	100	2079.20	1372.76	450	6393

Note. This table reports the distribution of votes for the Radical (UCR) and Conservative parties in the province of Buenos Aires in the elections of 1931, 1935, 1936, 1940, and 1941. The units of analysis are the electoral returns a the *partido* level. The partidos were the province's basic local administrative units. Sources: *Memorias* of the Ministerio del Interior and *Diario de Sesiones* of the Argentine Chamber of Deputies.

[13] We decided to exclude from the analysis the national elections that took place on November 8, 1931 and on March 4, 1934. Following a strategy of electoral abstention, the Radical party did not participate in these electoral contests. As such, these elections are not suited for the examination proposed in this paper. We also excluded from the analysis the 1938 legislative elections because we were unable to obtain disaggregated data.

It is tempting to conclude that Conservative victories entailed electoral fraud. The data, however, are insufficient to reach such conclusion. In other words, the uncertainty about the elections' class membership remains. Based on historical accounts, we do have a presumption regarding their values. Specifically, the 1935, 1936, and 1941 elections are often considered to be fraudulent, whereas the 1931 and 1940 contests are presumed clean. It is impossible, however, to be absolutely certain about the legitimacy these elections (i.e., we do not have an unambiguously "real" value—fraud/clean—for each of these five outcomes). The task is to evaluate the validity of the conventional wisdom using the vote counts themselves as our main source of evidence.

Our analysis will be based on the distribution of the digits in the vote counts at the partido level. The appropriateness of using this level of aggregation is highlighted by the data presented in Table 1. In each of the partidos, voters were served by a number of *mesas* (polling stations). These mesas were designed to include roughly the same number of voters (typically 122 voters). Yet, the number of voters per polling station varied widely, including a minimum of 86 (Ayacucho) and a maximum of 172 (Florencio Varela). Moreover, the number of mesas per partido also varied considerably. On average, each partido had 28.5 polling stations, with a minimum of 7 (General Guido and General Rodriguez) and a maximum of 78 (Pergamino). The uneven distribution of district sizes gives us confidence in the usefulness of focusing on the first significant digits (FSD) of the vote counts: Even if a party has roughly the same level of support in all the partidos, which means the party's share of the votes is roughly the same in all of them, then the vote counts will not necessarily have the same first digit in all the districts (Mebane 2008b).

2.2 *Synthetic Data*

As mentioned above, synthetic data can be designed to demonstrate certain key properties which may not be apparent in the authentic data. In order to assess the benefits of digital analysis as a useful fraud-detection tool, we generated a large amount of synthetic data with known levels of "contamination." The main advantage of this methodology is that, because we can specify the degree of manipulation in a controlled environment, it allows us to verify the sensitivity (i.e., detecting electoral manipulation cases when they were manipulated) and specificity (i.e., classifying clean elections when no manipulation was made) of our supervised learning model.

2.2.1 Data generation methodology

Synthetic data can be defined as data that are generated by simulated agents in a simulated system, performing simulated actions (Lundin, Kvarnström, and Jonsson 2002). We simulate a large number of electoral contests using Monte Carlo methods. A Monte Carlo experiment is analogous to a laboratory situation, where a real world scenario is replicated numerous times, and every time a different sample is drawn. These samples are then used to identify how electoral manipulation can lead to vote counts that do not satisfy Benford's Law. This law specifies that in a collection of numbers, the probability of the first or leading digit being d should be

$$P(\text{leading digit} = d) = \log_{10}\left(\frac{d+1}{d}\right),\ d \in \{1, 2, \ldots, 9\}. \tag{1}$$

Traditional analysis of Benford's Law considered its applicability to data sets (Varian 1972). In more recent scholarship, though, the focus has shifted to the study of *probability distributions* that obey Benford's Law. These studies demonstrate that, if applicable, Benford's Law is invariant under (1) an arbitrary change of scale; (2) an arbitrary raising to a power; and (3) an arbitrary change of the numerical basis (Hill 1995a, 1995b; Leemis, Schmeiser, and Evans 2000; Grendar, Judge, and Schechter 2007; Ciofalo 2009; Fewster 2009).[14] From a practical standpoint, Benford's Law is known to work better when the data in the sample cover several orders of magnitude and are not "artificially" biased in favor of any particular

[14] Analyses of distributions satisfying Benford's Law have also revealed that Benford compliance should not be expected in every random distribution (Leemis, Schmeiser, and Evans 2000). Yet, as Ciofalo (2009) points out, a sufficient condition for Benford's Law to hold is that the data in a given sample follow one of the following distributions: (1) a hyperbolic probability density function $p(d) \sim 1/d$; (2) a geometric progression (i.e., a sequence of numbers where each term after the first is found by multiplying the previous one by a fixed nonzero number); and (3) an exponential rank-size distribution $f(k) \sim \exp(-k/kn)$.

value. Thus, we typically expect Benford adherence in the case of large data sets whose numbers are the result of a mathematical combination of distributions and where the mean is greater than the median with a positive skew (Cho and Gaines 2007; Janvresse and de la Rue 2004).[15]

We wrote an R function to implement our simulations (see Appendix 1). The function generates vote counts for 100 simulated districts, each containing two competing parties, $i \in \{A, B\}$. The total number of votes for party i in district $j = \{1, \ldots, 100\}$ is determined by:

$$V_{ij} = \alpha_i X_i, \tag{2}$$

where α_i is a constant and X_i denotes a random variable with Benford's distribution.

Following Leemis, Schmeiser, and Evans (2000), we generate the Benford variate X by

$$X \leftarrow \lfloor 10^U \rfloor, \tag{3}$$

where $U \sim U(0, 1)$.

Notice that V_{ij} is composed by both deterministic and stochastic elements. The first one, α_i, can be interpreted as the baseline electoral support for party i across all districts. Hence, this parameter captures the idea that "a rising tide lifts all boats" (i.e., an electoral landslide). Namely, if say party A becomes more popular that party B in a particular election (because it files more attractive candidates or any other reason), then $\alpha_A > \alpha_B$. The second element, the random variable X_i, does not depend on the baseline popularity of the parties. It can thus be interpreted as idiosyncratic variance in the electoral support for party i.[16] In addition, estimating V_{ij} in the proposed way allows for the possibility that the party with the lower α may still be the winner in some of the districts (i.e., if $\alpha_A > \alpha_B$, party B would win in those districts j, where $X_{Bj} > \frac{X_{Aj}\alpha_A}{\alpha_B}$).

The historical evidence suggests that acts of fraud were often committed in almost every polling station throughout Buenos Aires, and they usually consisted of taking votes away from the Radical opposition and giving them to the Conservatives. To capture these specific characteristics of electoral fraud, we simulate electoral tampering in the following way: in every district $j = \{1, \ldots, 100\}$, we take away a fixed proportion $\gamma > 0$ of party A's votes and give $\delta(\gamma V_{Aj})$ votes to party B (with $\delta > 0$). Therefore, whenever fraud is committed, the total number of votes for each of the parties in every district j becomes $V_{Aj} = (1 - \gamma)(\alpha_A X_{Aj})$ and $V_{Bj} = (\alpha_B X_{Bj}) + \delta[\gamma(\alpha_A X_{Aj})]$, respectively.

So, for example, consider $\alpha_A = 400$ and $\alpha_B = 320$ and suppose that in district $j = 1$, $X_{A1} = X_{B1} = 2$. The total number of votes for parties A and B in district 1 would be given by $V_{A1} = (400 \times 2) = 800$ and $V_{B1} = (320 \times 2) = 640$, respectively. Without fraud, party A would carry the district. However, if $\gamma = 0.3$ and $\delta = 1.2$, party A's vote count would be given by $V_{A1} = (0.7) \times (800) = 560$, whereas party B would obtain $V_{B1} = (640) + (1.2)[(0.3) \times 800] = 928$ votes, reversing the outcome.

Using our R function, we generated 10,000 electoral contests. To create different types of elections, we treated each contest as a Bernoulli trial with two possible outcomes. We set $p = .5$ as the probability of success/failure to obtain balanced class sizes. Fraudulent elections were deliberately manipulated by taking votes away from party A and giving them to party B using values of $\delta > 0$ and $\gamma > 0$. In the case of clean elections, the parameters δ and γ were set to zero.

2.2.2 Data description

The uneven distribution of district sizes in Buenos Aires exhibits the kind of complexity that can produce counts with first digits that follow Benford's Law. As such, following Busta and Weinberg (1998) and Grendar, Judge, and Schechter (2007), we focus on two variables based on the FSD of the simulated vote counts: the first moment of the FSD distribution and the frequency of the number one as the FSD. Specifically, in each of our simulated elections, we consider (1) the mean of the first digit of party i's votes

[15] Several recent studies use Benford's Law to identify electoral tampering (Mebane 2006, 2008b; Beber and Scacco 2008). In this respect, our approach is most similar to recent work by Mebane (2007). But, in contrast to Mebane, who calibrates his diagnostic device using the very same data he seeks to diagnose, we employ the synthetic data as our training set and use authentic data as our test set.

[16] In expectation, the value of X should be 3.162 (the square root of 10), for both parties.

in every district, $\bar{d}^i = \sum_{k=1}^{9} d_k \frac{\sum_{j=1}^{100} d_j=k}{100}$ and (2) the frequency of the number 1 as the FSD of party i's votes in every district, $p(d^i = 1) = \frac{\sum_{j=1}^{100} d_k=1}{100}$.

So, for instance, consider an election with $j = 100$ districts where the number 1 appears as the first digit for party A's votes in 12 districts and the other 8 digits appear as the first digit for party A's votes in exactly 11 districts each. The mean of the FSD can be calculated as the following: $\bar{d} = (1 \times \frac{12}{100}) + (2 \times \frac{11}{100}) + (3 \times \frac{11}{100}) +, \cdots, +(9 \times \frac{11}{100}) = 4.96$, whereas the frequency of the first digit is $p(d = 1) = \frac{12}{100} = 0.12$.

For each of the 10,000 simulated electoral contests, we recorded the values of our variables of interest, $\bar{d}^i$ and $p(d^i = 1)$. To distinguish between legitimate and fraudulent elections, we also recorded the outcome of each Bernoulli trial. Therefore, our training set includes the following pieces of information for each simulated election: a set of input objects (the variables $\bar{d}^1$ and $p(d^i = 1)$) and an output object (a dummy variable indicating whether the vote counts were manipulated or not).

2.3 *Calibration*

The process of applying supervised machine learning to an actual problem requires that the training set possesses characteristics of the real-world relationship between our input and output variables. To ensure that the synthetic data are representative of the authentic data, we calibrate our simulations using the first two elections in our data set (1931 and 1935).

The main goal is to find values of α_A and α_B and the *fraud* parameters, γ and δ, that will provide the best fit between the simulated and the real data. We proceed in the following way: first, we generate a simulated electoral contest using the methodology laid out in the previous section. Next, we record the vote count for our two parties in each of the 100 electoral districts, with and without fraud. So, for example, in the simulation, party A may win district number 36 by a vote of 1884 to 748 in a clean contest, but under fraud, party B (the beneficiary of electoral manipulation) may carry the same partido by a count of 1426 to 1319. More generally, our simulation yields four vote distributions: a nonfraudulent vote count for party A, a fraudulent vote count for party A, a nonfraudulent vote count for party B and a fraudulent vote count for party B. Per the characterization of the fraudulent practices during the "infamous decade" discussed above, we compare the electoral returns of party A to those of the Radicals (the victims of fraud) and party B's vote count to that of the Conservatives (the main beneficiaries of electoral tampering).

To carry out our comparisons, we use a two-sample Kolmogorov–Smirnov (K-S) test. This test is one of the most useful and general nonparametric methods for comparing two samples. Table 2 reports the D statistic and associated p values of the K-S tests for each of our pair of distributions under different parameter values.[17] Entries where a statistically significant difference (at the 99% confidence level) between the groups exists are indicated in bold.

The results presented in Table 2 suggest that $\alpha_A = 400$, $\alpha_B = 320$, $\gamma = 0.3$, and $\delta = 1.2$ provide the best fit between the simulated and the real data. In addition, the second panel of Table 2 shows that while the vote distributions for party B seem to match the Conservative's vote counts, a statistically significant difference between the vote distributions for party A and the Radical's vote counts does exist. This finding indicates that manipulation of the electoral process, rather than a shift in voters' preferences, account for the electoral outcomes (i.e., the Conservatives' gains do not match the Radicals' losses).

To further demonstrate the validity of our calibration parameters, Fig. 2 presents a graphical comparison between pairs of distributions. Consider the nonfraudulent vote count for party A and the Radical party's electoral returns in the 1931 elections. As the graph for this election demonstrates, the two distributions are strikingly similar. Likewise, the fraudulent vote count for party B and the Conservative's electoral returns in the 1935 election are also almost identical.

The analysis of the vote counts presented in this section illustrates how traditional statistical tests can be used to assess the validity of a Monte Carlo–generated training set. Beyond this practical demonstration, the results provide an important validation for this approach, as the simulated vote distributions under

[17] The D statistic quantifies a distance between the empirical distribution functions of two samples, and its distribution is calculated under the null hypothesis that the samples are drawn from the same distribution.

Table 2 Comparison of authentic and synthetic data
Kolmogorov–Smirnov test: D statistic (p values)

	Radical party		*Conservative party*	
Election	*Clean*	*Fraud*	*Clean*	*Fraud*
		$\alpha_A = 400, \alpha_B = 320, \delta = 0, \gamma = 0$		
1931	**0.23** (0.008)		**0.24** (0.005)	
1935	**0.23** (0.008)		**0.32** (0.000)	
		$\alpha_A = 320, \alpha_B = 520, \delta = 0, \gamma = 0$		
1931	**0.28** (0.001)		0.18 (0.069)	
1935	**0.24** (0.005)		0.21 (0.020)	
		$\alpha_A = 400, \alpha_B = 320, \delta = .1, \gamma = .6$		
1931	0.15 (0.193)	0.21 (0.020)	**0.29** (0.000)	**0.23** (0.008)
1935	**0.30** (0.000)	**0.27** (0.001)	**0.37** (0.000)	**0.33** (0.000)
		$\alpha_A = 400, \alpha_B = 320, \delta = .3, \gamma = 1.2$		
1931	0.14 (0.281)	**0.32** (0.000)	0.21 (0.024)	**0.27** (0.001)
1935	**0.39** (0.000)	**0.24** (0.006)	**0.34** (0.000)	0.1 (0.699)
		$\alpha_A = 400, \alpha_B = 320, \delta = .5, \gamma = 1.8$		
1931	0.19 (0.047)	**0.49** (0.000)	**0.23** (0.008)	**0.50** (0.000)
1935	**0.28** (0.001)	0.16 (0.140)	**0.38** (0.000)	**0.23** (0.008)

Note. This table reports the comparison between the authentic and synthetic data. To carry out our comparisons, we use a two-sample Kolmogorov–Smirnov (K-S) test. Entries in bold indicate statistical significance at a 99% level. The results suggest that $\alpha_A = 400$, $\alpha_B = 320$, $\gamma = .3$, and $\delta = 1.2$ provide the best fit between the simulated and the real data.

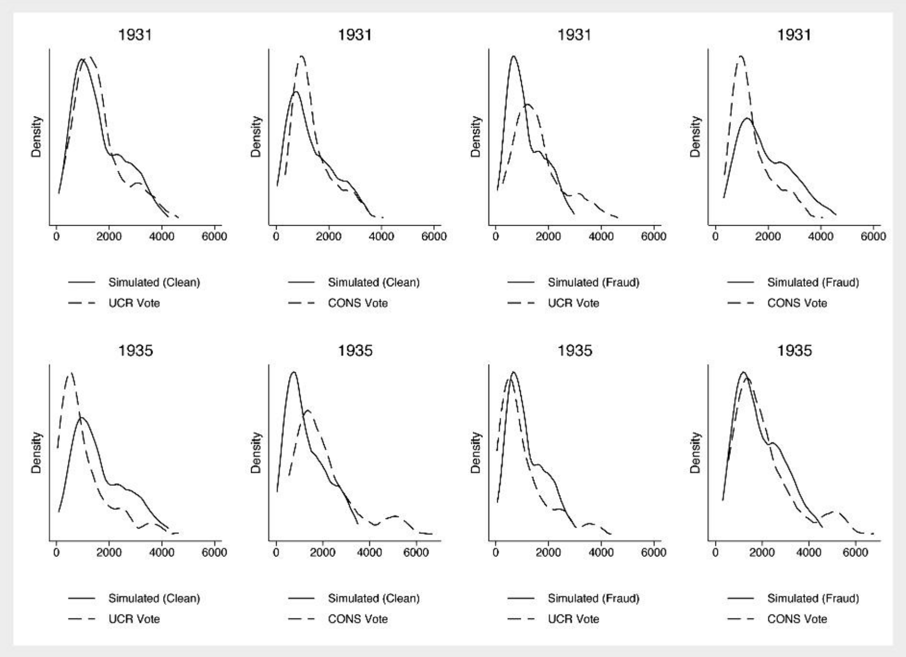


Fig. 2 Comparison of authentic and synthetic data. This figure presents a graphical comparison between pairs of distributions: the fraudulent/nonfraudulent vote count for parties A and B in our simulations and the actual electoral returns for the Radicals and the Conservatives in the 1931 and 1935 elections. As the graph for the 1931 election demonstrates, the nonfraudulent distribution of the vote counts for party A and the electoral returns for the Radical party are strikingly similar. Likewise, the fraudulent vote count for party B and the Conservative's electoral returns in the 1935 election are also almost identical.

the two different scenarios of interest (fraud/not fraud) bear a close resemblance to the actual data. They also reflect the conventional wisdom regarding these electoral contests. We now turn our attention to algorithm selection, the next step in the application of supervised machine learning to our particular problem.

3 Learning Algorithm

As stated above, the goal of a learning algorithm is to construct a classifier given a set of training examples with class labels. Assuming that only two classes exist, then the problem can be stated as follows: produce a classifier that distinguishes between examples from two classes (labeled as $y = 0$ and $y = 1$) on the basis of m observed predictor variables (also known as features) $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T \in \mathbf{R}^m$.

Specifically, given a training set $D = \{(\mathbf{x}^{(i)}, y^{(i)} : \mathbf{x}^{(i)} \in \mathbf{R}^m, y^{(i)} \in \{0, 1\}\}_{i=1}^n$ with n labeled examples (where $y^{(i)}$ is the label associated with example $\mathbf{x}^{(i)}$), the two principal tasks are to (1) identify a subset of the features that is most informative about the class distinction (feature selection) and (2) learn a function that most accurately predicts the class of a new example (classifier design). So, for example, if the problem is identifying fraudulent elections, then $\mathbf{x}^{(i)}$ is some representation of a given electoral contest and $y^{(i)} \in$ {"Clean", "Fraudulent"}. We start first with task number two, and we address the former (feature selection) in Section 4.1.

3.1 *Classification Using Naive Bayes*

The classification problem can be written as the problem of finding the class with maximum probability given a set of observed attribute values. Such probability is seen as the posterior probability of the class given the data and is usually computed using Bayes' theorem. Therefore, the probability of an example $\mathbf{x}$ being class $y \in \{0, 1\}$ is

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}. \tag{4}$$

Example $\mathbf{x}$ is classified as the class $y = 1$ if and only if

$$\log \frac{p(y = 1|\mathbf{x})}{p(y = 0|\mathbf{x})} \geqslant 1. \tag{5}$$

Assume that all attributes are independent from each other given the class; namely,

$$p(\mathbf{x}|y) = p(x_1, x_2, \ldots, x_m|y) = \prod_{i=1}^{m} p(x_i|y). \tag{6}$$

This allows us to write, following Bayes' theorem, the posterior probability of the class y as:

$$p(y|\mathbf{x}) = \frac{p(y)}{p(\mathbf{x})} \prod_{i=1}^{m} p(x_i|y). \tag{7}$$

As shown above, a probability ratio $g(\mathbf{x}) = \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})}$ can be expressed in terms of a series of likelihood ratios, such that example $\mathbf{x}$ is classified as the class $y = 1$ if and only if $g(\mathbf{x}) \geqslant 1$.

The function $g(\mathbf{x})$ is usually called a naive Bayes classifier (NB), or "idiot's Bayes," and is one of the simplest classification algorithms. Its main source of simplicity is the independence assumption. As made clear in equation 6, naive Bayes assumes that the features in the problem are independent. This assumption largely simplifies the learning process, as the classifier is fully defined simply by the conditional probabilities of each attribute given the class (Zhang 2004). Hence, Bayes' methods train very quickly since they require only a single pass on the data either to count frequencies (for discrete variables) or to compute the normal probability density function (for continuous variables under normality assumptions).

When the strong independence assumption is satisfied, NB is optimal, that is, guarantees minimum classification error (Zhang 2004; Kuncheva 2006). This assumption is rarely true in most real-world applications. Numerous experimental studies, however, demonstrate that NB is accurate and efficient even if the independence assumption is violated. Zhang (2004) shows that regardless of how strong the dependences among attributes are, NB can still be optimal if the dependences distribute evenly in classes or if the dependences cancel each other out. Kuncheva (2006) demonstrates that, given two binary features and two equiprobable classes, NB is optimal for dependent features as long as the covariances for the two classes are equal. Moreover, Rish (2001); Demirekler and Altınçay (2002) and Altınçay (2005) show that dependencies may be good for improving the accuracies provided by the NB classifier.

The robust performance of NB has been attributed to various estimation properties (Domingos and Pazzani 1997; Hand and Yu 2001). For example, Domingos and Pazzani (1997) perform a large-scale comparison of NB with algorithms for decision tree induction, instance-based learning, and rule induction on a series of data sets. They find NB to be often superior to the other learning schemes, even on data sets with substantial feature dependencies. The most important explanation for NB's success, however, lies in the fact that conditional independence is only a sufficient but not a necessary condition for optimality (Domingos and Pazzani 1997; Hand and Yu 2001; Rish 2001; Zhang 2004; Zhang and Su 2004). From a classification point of view, the relative values of a posteriori probabilities assigned to different hypotheses are more important than the accuracy of their estimates. Indeed, the accuracy of approximation of the likelihoods of joint observations, $p(x_1, x_2, \ldots, x_m|y)$, is irrelevant as long as for any example $\mathbf{x}$, the largest posterior probability corresponds to the same class as with the true posterior probabilities (Hastie, Tibshirani, and Friedman 2009). Therefore, despite its naive design and apparently oversimplified assumptions, naive Bayes classifiers often outperform far more sophisticated alternatives, even if the independence assumption does not hold.

3.2 *Estimation*

From an estimation viewpoint, the classification problem can be considered as one in which the goal is to estimate a function of the form $P(\text{class}|\mathbf{x}) = f(\mathbf{x})$. All model parameters (i.e., class priors and feature probability distributions) can be approximated with relative frequencies from the training set. In the presence of discrete and Gaussian data, this process turns out to be straightforward.

For our two-class problem, we can fit density estimates $\hat{f}_y(\mathbf{x})$, $y \in \{0, 1\}$ separately in each of the classes, and we also have estimates of the class priors $\hat{\pi}_y$ (the sample proportions). Then

$$p(y = 1|\mathbf{x}) = \frac{\hat{\pi}_1 \hat{f}_1(\mathbf{x})}{\hat{\pi}_1 \hat{f}_1(\mathbf{x}) + \hat{\pi}_0 \hat{f}_0(\mathbf{x}).} \tag{8}$$

As noted above, naive Bayes assumes that given a class y, the features x_i are independent:

$$f_y(\mathbf{x}) = \prod_{i=1}^{m} f_{yi}(x_i). \tag{9}$$

Hence, when the components x_i of $\mathbf{x}$ are discrete, an appropriate histogram estimate can be used. This provides a very simple way of mixing variable types in a feature vector (Hastie, Tibshirani, and Friedman 2009, 211). But, if we are dealing with continuous variables, then the domain of attributes needs to be partitioned. A common way to handle continuous attributes in NB classification is to use Gaussian distributions to represent the likelihoods of the features conditioned on the classes (Mitchell 1997; Bustamante, Garrido, and Soto 2006; Hastie, Tibshirani, and Friedman 2009). Namely, the individual class-conditional marginal densities f_{yi} can each be estimated separately using one-dimensional kernel density estimates.

Starting from equation 8, the logit transform can be derived as:

$$\log\frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \log\frac{\pi_1 f_1(\mathbf{x})}{\pi_0 f_0(\mathbf{x})} = \log\frac{\pi_1 \prod_{i=1}^{m} f_{1i}(x_i)}{\pi_0 \prod_{i=1}^{m} f_{0i}(x_i)}$$

$$= \log\frac{\pi_1}{\pi_0} + \sum_{i=1}^{m}\log\frac{f_{1i}(x_i)}{f_{0i}(x_i)} = \alpha_1 + \sum_{i=1}^{m} g_{1i}(x_i). \tag{10}$$

This has the form of a *generalized additive model* (Hastie, Tibshirani, and Friedman 2009, 211). The model can be fitted in different ways; however, in the case of two mutually exclusive classes, as it is the case here, the conversion of the log-likelihood ratio to a probability takes the form of a sigmoid curve.

4 Empirical Analysis

4.1 *Feature Selection*

As discussed above, an important factor that must be taken into account when choosing a learning algorithm is the nature of the data to be classified. Generally, statistical learning systems tend to perform better when a few meaningful features are used. By reducing the dimensionality of the data (i.e., avoiding the "curse of dimensionality"), data mining algorithms tend to operate faster and more effectively. Feature subset selection is thus an important step in the evaluation of a learning algorithm. This process consists of identifying and removing as many irrelevant and redundant features as possible (Yu and Liu 2004; Kotsiantis 2007).

A classifier's evaluation is often based on its prediction accuracy. Four measures are usually used to assess a classifier's performance. The most popular of these indicators, prediction accuracy, measures the proportion of correctly classified instances. The other three measures are positive predictive accuracy (the reliability of positive predictions on the induced classifier), sensitivity (the fraction of actual positive examples that are correctly classified), and specificity (the fraction of actual negative examples that are correctly classified) (Tan and Gilbert 2003; Hastie, Tibshirani, and Friedman 2009).

For our two-class classification rule, recall the logit transform discussed in the previous section. We relate the mean of the binary response $\mu(\mathbf{x}) = p(y = 1|\mathbf{x})$ to the predictors via a linear regression model and the *logit* link function:

$$\log\left(\frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})}\right) = \alpha + \beta_1 x_1 +, \cdots, +\beta_m x_m. \tag{11}$$

We evaluate the classification rule using all the data in our training set (10,000 observations). Per the calibration results presented above, we use $\alpha_A = 400$ and $\alpha_B = 320$, and we set $\gamma = \delta = 0$ when we simulate a clean election and $\gamma = 0.3$, $\delta = 1.2$ otherwise. As expected, these parameter values minimize our classification errors (see Fig. A2).[18]

The dependent variable, *Fraud*, takes the value of one if an observation corresponds to a simulated electoral contest in which the data were purposively "contaminated" by electoral tampering and zero otherwise. Recall that we simulated these different types of elections by treating each contest as a Bernoulli trial with two possible outcomes and set probability of success/failure as $p = .5$. Therefore, by design, in roughly half of our observations, the variable *Fraud* takes the value of one and in the other half, a value of zero.

To avoid overfitting, we only use the first moment of the FSD distribution, and the frequency of the number one as the FSD corresponding to the simulated vote counts of the party benefitting from fraud

[18] As Fig. A2 demonstrates, the classifier's performance regarding the training data does not depend on these choices: even parameter values reflecting different levels of fraud would yield very similar results. The optimal parameter values, however, ensure that we obtain appropriate class-conditional marginal densities. The results presented in Table A2 show what happens when incorrect parameter values are employed. When the simulated amount of fraud is scant ($\gamma = .1$, $\delta = .6$), the classifier's performance regarding the Buenos Aires elections suffers significantly: all elections are considered clean (with a correct classification rate of 50%). Likewise, if we simulate too much fraud ($\gamma = .5$, $\delta = 1.8$), the classifier also performs quite badly, but in the opposite direction: all elections are considered fraudulent (with a correct classification rate of 50%).

(i.e., $\bar{d}^B$ and $p(d^B = 1)$) as our predictor variables. Another reason why we exclude the other party's vote count from the analysis is the following: recall that whenever fraud is committed, the total number of votes for the party affected by fraud in each district is $V_{Aj} = (1-\gamma)(\alpha_A X_{Aj})$. That is, the postmanipulation vote count for the party affected by fraud is some fraction of its original vote. As discussed above, Benford's Law is scale invariant. Hence, this additional information would be of little help to ascertain whether electoral manipulation leads to vote counts that do not satisfy Benford's Law.

Given that both features are based on the FSD distribution of party B's votes, the independence assumption is clearly violated. Nonetheless, the dependency of the feature does not vary across our two categories (the correlation coefficients between the two variables under each class label are $\rho = -0.63$ and $\rho = -0.64$ in the clean and manipulated elections, respectively). As discussed above, naive Bayes is an effective classification tool when the dependencies among the features are independent of the class labels (Zhang 2004). Therefore, despite the strength of dependence between our two features, we can still take advantage of the naive Bayes classifier as a learning scheme.

Table 3 shows the test error rates. Using this contingency table, we can evaluate the classifier's performance. The overall correct classification rate is 94.28%. The positive predictive accuracy is 94.28%, the sensitivity is 94.13%, and the specificity is 94.43%. In the case of electoral fraud, the usual concern is to avoid false negatives, that is, we want to avoid classifying a fraudulent election as a clean one. As Table 3 indicates, the probability that the classifier would produce a false negative is very low (5.87%).

A receiver operating characteristic (ROC) curve is a two-dimensional visualization of the tradeoff between true- and false-positive rates of a classification algorithm. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The ROC plot is given in Fig. 3.

The area under the ROC (AUROC) is a single-figure summary measure associated with ROC performance assessment (range [0, 1];1 being optimal). This measure can be interpreted as the probability that when one positive and one negative examples are randomly selected, the classifier will assign a higher score to the positive example than to the negative. As Fig. 3 demonstrates, the proposed model (i.e., input representation) provides an outstanding classification rule, with an AUROC of 0.98.

4.2 *Classification of Buenos Aires' Elections*

Having demonstrated the accuracy of our classification rule, we now turn to the evaluation of the classifier's performance on the authentic data. Since we have very few authentic fraudulent elections, we use here some of the same information employed as seed data for the synthetic data generation process. Namely, we rely on the 1931 and 1935 elections. In the next section, we use the other three electoral contests in our data set to further validate our classification results.[19]

Figure 4 presents class-conditional density estimates obtained from our training data. Each of the two panels on the left show two separate density estimates for our two predictor variables in the fraudulent versus nonfraudulent groups, using a Gaussian kernel density estimate in each. And each of the two panels on the right show the normal cumulative density function for the two predictor variables, once again, in the fraudulent versus nonfraudulent groups.

Table 3 Contingency table: logit model fit to the training data

	Predicted class	
True class	*Clean (0)*	*Fraud (1)*
Clean (0)	4777	282
Fraud (1)	290	4651

Note. This table reports the classifier's performance regarding the training data. The overall correct classification rate is 94.28%. The positive predictive accuracy is 94.28%, the sensitivity is 94.13%, and the specificity is 94.43%.

[19] For a similar approach, see Lundin, Kvarnström, and Jonsson (2002).

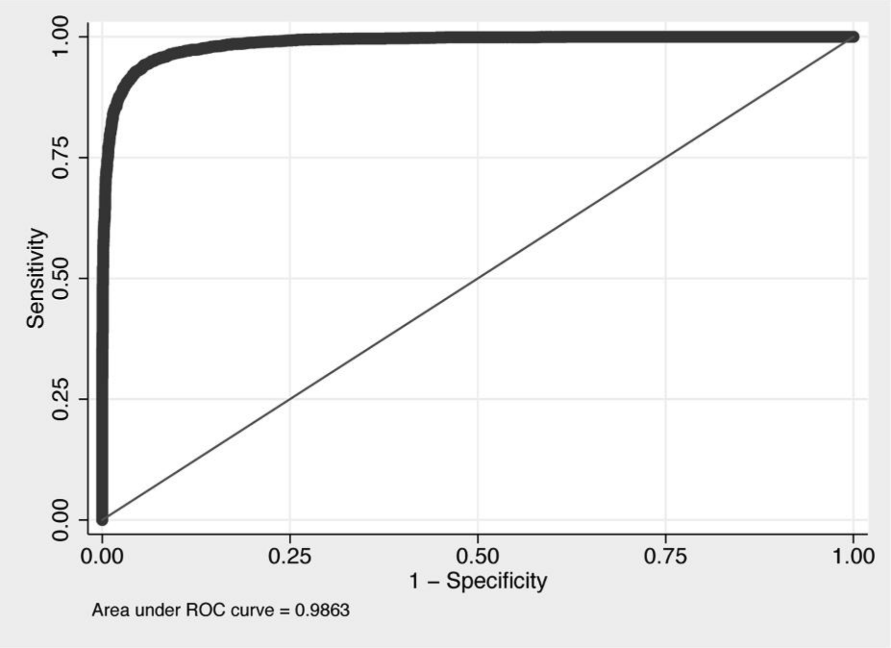


Fig. 3 ROC curve for classification rules fit to training set. This figure presents the ROC plot of our classifier's performance. It is a two-dimensional visualization of the tradeoff between true- and false-positive rates of a classification algorithm. As the figure demonstrates, our proposed model provides an outstanding classification rule, with an AUROC of 0.98.

The evidence in Fig. 4 suggests that we can effectively use the class-conditional marginal densities for classification purposes. More specifically, we can use the distribution function of our two predictor variables and combine them with probabilities of the naive Bayes model in a straightforward manner. The procedure is the following: we are interested not in the probability that the value of our feature variable x_i is a particular number but rather in the probability that x_i has a value *less than or equal to* some critical number. So, suppose that the random variable x_i is the mean of the FSD and that we are interested in the event that x_i is less than or equal to 3. Given the class-conditional distributions presented in Fig. 4, we can calculate $F(3) = p(x_i \leqslant 3)$ for both the fraudulent and nonfraudulent groups. These are $F_1(3) = p(x_i \leqslant 3|y = 1) = 0.82$ and $F_0(3) = p(x_i \leqslant 3|y = 0) = 0.0162$, respectively.

We introduce now our validation set, composed by electoral data from the province of Buenos Aires. As discussed above, the historical evidence suggests that the Conservative party was the beneficiary of fraudulent practices throughout this period. Hence, for each election, we can calculate our two feature variables based on the FSD distribution of the votes of the Conservative party. For instance, in the 1931 elections, the frequency of the number 1 is 0.4 and the mean of the FSD is 3.77.

Recall from equation 7 that, following Bayes' theorem, we can calculate the posterior probability of class y as $p(y|\mathbf{x}) = \frac{p(y)}{p(\mathbf{x})} \prod_{i=1}^{m} p(x_i|y)$. Therefore, we can use the FSD information to classify each electoral contest. To illustrate how the classifier works, let us focus on the 1931 elections. The probability that the election was clean can be written as:

$$p(c|\mathbf{x}) = \frac{p(c)p(x_1 \leqslant 0.4|c)p(x_2 \leqslant 3.77|c)}{p(c)p(x_1 \leqslant 0.4|c)p(x_2 \leqslant 3.77|c) + p(c')p(x_1 \leqslant 0.4|c')p(x_2 \leqslant 3.77|c')},$$

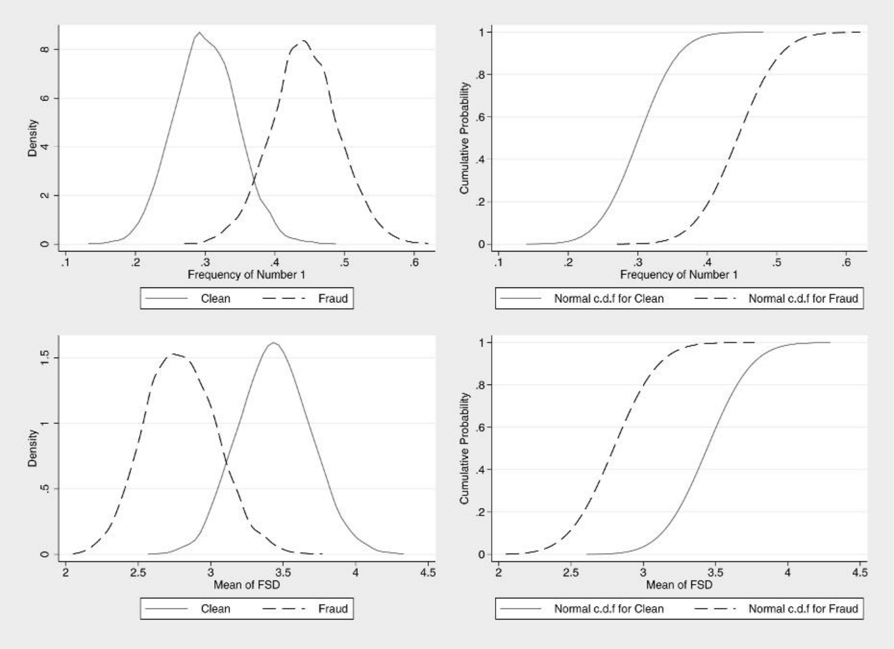


Fig. 4 Class-conditional density estimates. This figure presents class-conditional density estimates obtained from our training data. Each of the two panels on the left show two separate density estimates for our two predictor variables in the fraudulent versus nonfraudulent groups, using a Gaussian kernel density estimate in each. And each of the two panels on the right show the normal cumulative density function for the two predictor variables, once again, in the fraudulent versus nonfraudulent groups. The evidence suggests that we can effectively use the class-conditional marginal densities for classification purposes.

where $p(c) = p(\text{clean})$, $p(c') = p(\text{fraud})$, x_1 denotes the frequency of the number 1, and x_2 denotes the mean of the FSD. Given a prior assignment of probabilities $p(c) = p(c') = \frac{1}{2}$, the updated probability that the 1931 election was clean is equal to

$$p(c|\mathbf{x}) = \frac{(.5)(1)(.9108)}{(.5)(1)(.9108) + (.5)(.1954)(1)} \approx 0.823.$$

The top panel of Table 4 presents the classification of the 1931 and 1935 elections obtained using the NB learning algorithm. Our results corroborate the validity of our approach. The labels assigned by the classifier, despite using uninformative priors (i.e., $p(\text{clean}) = p(\text{fraud}) = \frac{1}{2}$) and a minimal amount of information, allow us to discriminate between the two types of electoral contests.[20]

For instance, take the April 5, 1931 provincial election. This was the first electoral contest that took place after the 1930 military coup. When all the votes were counted, the Radicals had 218,283 tallies to 187,734 for the Conservatives. The results shook the provisional military government to its foundation (Potash 1969). Citing irregularities in the voter registries, president Uriburu annulled the elections a few months later. Yet, the government's decision to annul the elections seemed to be completely unfounded. As Walter (1985) notes, the election was held in good order, with a few minor incidents. In fact, competing parties agreed that it was a free, open, and honest contest. In line with these historical accounts, the data presented in Table 4 indicate that the 1931 elections were unlikely to be fraudulent.

In contrast with the 1931 election, the November 3, 1935 gubernatorial election was immediately and universally condemned as one of the most fraudulent and irregular in Argentine history (Bejar 2005). The

[20] We evaluate the sensitivity of these results to different assumptions regarding the prior probabilities of observing a fraudulent/clean election in Fig. A3. The analysis indicates that our results are robust to different expectations. In the case of fraudulent elections, correctly classified outcomes entail posterior probabilities higher than .5, for the whole range of priors and for clean elections for prior probabilities, $p(\text{clean}) > .25$.

Table 4 Classification of Buenos Aires' elections (1931–1941)

Election	$p(clean) = p(fraud)$	$p(clean\|\mathbf{x})$	$p(fraud\|\mathbf{x})$	$log\, \frac{p(y=1\|\mathbf{x})}{p(y=0\|\mathbf{x})}$	*Classification*
		Validation set (seed data)			
1931	0.5	0.823	0.176	−1.539	Clean
1935	0.5	0.054	0.945	2.845	Fraudulent
		Test Set			
1940	0.5	0.756	0.243	−1.135	Clean
1941	0.5	0.080	0.919	2.441	Fraudulent

Note. This table reports the classification of the elections in our validation set (top panel) and in our test set (bottom panel) obtained using the NB learning algorithm.

results underscored the extent of the irregularities. The Conservative slate defeated the Radical candidates by more than 100,000 votes, 278,533 to 171,081. As Table 4 shows, our *learner* unambiguously classified this election as fraudulent.

4.2.1 Evaluation of the Classifier's performance

From a classification point of view, the results presented in the top panel of Table 4 demonstrate that our *learner* can be effectively used as a tool for identifying fraud. The last step is to evaluate the performance of the classifier vis-a-vis the conventional wisdom. For this purpose, we turn now our attention to the 1940 and 1941 electoral contests.

The March 3, 1940 legislative elections took place under a peculiar political climate. Incumbent president Roberto Ortiz had pledged, in his opening address to the the National Congress in 1938, to end fraudulent political practices and to restore democracy. Under the threat of a federal intervention, local authorities refrained from engaging in electoral fraud. On April 19, 1940, the *Review of the River Plate* praised "… the correctitude with which the voting was conducted …". Indeed, the elections were the "… freest and most democratic since April 1931 …" according to Walter (1985, 178). These accounts, once again, corroborate the validity of our approach: The findings presented in the bottom panel of Table 4 clearly suggest that this election was legitimate.

The last provincial election of the "infamous decade" took place on December 7, 1941. The Conservative gubernatorial candidate won by almost 100,000 votes. Unlike the elections from the previous year, Ortiz was no longer in charge. Due to medical problems, he had to give up his powers to vice president Castillo in July of 1940. As Alston and Gallo (2010) note, the election was full of irregularities (public voting, police harassment, ballot box stuffing). Indeed, this election is identified as fraudulent by our detection tool.

To further validate our results, we focus now on the national congressional elections held on March 1, 1936. This is a particularly valuable set because (1) we did not use these elections as seed data for the synthetic data generation process and (2) official records documenting electoral irregularities do exist (and, therefore, we have a real value of fraud/clean for our output variables).

Originally scheduled for late 1935, these elections were as fraudulent as the gubernatorial election held in that year. In fact as Walter points out, on March 8, the national election board "… annulled the results of 259 polling stations in 72 districts involving 63,000 voters and convoked complementary elections for March 15 …" (Walter 1985, 156). Using the official records of the Argentine Chamber of Deputies (cf. Cámara de Diputados. *Diario de Sesiones*, June 17, 1936, 942–943), we identified those districts where irregularities acknowledged by the election board took place. After removing the ten partidos that were excluded from our previous analysis, our sample for the 1936 elections includes 66 districts with faulty polling stations and 34 where no irregularities were uncovered. Any district in which the results of at least one single polling station was annulled is thus considered a fraudulent district. If anything, using this criterion plausibly leads us to underestimate the detection capabilities of our classification tool.[21]

[21] The fraction of polling stations with irregularities in each of these fraudulent districts ranges from 1 of 48 (2.08%) in Lincoln to 7 of 21(33.33%) in Puan.

As before, we use our two feature variables based on the FSD distribution of the votes of the Conservative party for classification purposes. For the fraudulent districts, the frequency of the number 1 is 0.39 and the mean of the FSD is 3.12. In those districts where no irregularities were uncovered, the frequency of the number 1 is 0.44 and the mean of the FSD is 3.61. Given a prior assignment of probabilities $p(\text{clean}) = p(\text{fraud}) = 0.5$, the updated probability that the elections in those districts where irregularities were acknowledged by the election board were fraudulent is equal to $p(\text{fraud}|\mathbf{x}) \approx 0.613$. And the updated probability that the elections in those districts where no irregularities were uncovered was clean is equal to $p(\text{clean}|\mathbf{x}) \approx 0.596$.

4.3 *Standard Fraud Detection Algorithms*

Our findings clearly indicate that the method proposed in this paper can be effectively used to assess the fraudulent/nonfraudulent status of elections. To further demonstrate the advantages of our approach, it seems appropriate to explicitly show how standard vote fraud detection algorithms yield different and inferior results. In this section, we discuss some commonly used election forensics tools and apply them to the four elections in the province of Buenos Aires analyzed above.

Statistical analyses of electoral irregularities fall into two categories. The first looks for anomalies in the patterns of the numbers—the digits—employed in official protocols. The second examines anomalies in the distribution of turnout (Levin et al. 2009). With respect to the first category, the first-digit Benford's Law test (1BL) is often used as a fraud detection technique. This method employs a chi-square goodness-of-fit test to establish conformity with Benford's Law. More specifically, let $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$, where O_i and E_i are the observed and expected frequencies for digit i as the FSD, respectively. The null hypothesis is that the data follow the Benford distribution. The test statistic follows a χ^2 distribution, so the null hypothesis is rejected if $\chi^2 > \chi^2_{\alpha,8}$, where α is the level of significance (Cho and Gaines 2007).

In a similar vein, Mebane (2006, 2008b, 2010) looks for deviations from the "second-digit Benford's Law." These tests come in two forms. The first one, the second-digit Benford's Law test (2BL) is analogous to the 1BL test. The only difference is that it focuses on the relative frequency of the second significant digit (SSD).[22] The second test examines the first moment of the SSD distribution. If the vote counts' second digits follow Benford's Law, then the expected value for the second-digit mean is 4.187. Statistically significant deviations from this value are considered as evidence of electoral irregularities (Mebane 2006, 2008b, 2010).

An additional digit-based test focuses on the distribution of last digits. According to Beber and Scacco (2008), fair election procedures should produce returns where last digits occur with equal frequency. Hence, they propose a test based on the distribution of last digits. Specifically, their test focuses on the relative frequency of the last significant digit: an election is considered to be suspicious if the distribution of last digits departs significantly from what we would expect from a purely random process.

Turning to the second category of forensic indicators, Mebane (2008b) proposes a simple test to check if unusual turnout is associated with unusual vote totals. Specifically, for vote counts y_i observed for districts indexed by i, he regresses (by ordinary least squares) the SSD on turnout. The estimated coefficient for the turnout indicator when elections are nonfraudulent should be statistically undistinguishable from zero (Mebane 2008b).[23]

Finally, Levin et al. (2009) discuss a second indicator based on turnout rates. If artificially inflated turnout is absent, then the relationship between turnout (T) and a party's share of the eligible electorate (V/E) should approximately match the party's overall share of the vote. Therefore, regression estimates of the relationship between V/E and T in otherwise homogeneous data should fail in the interval [0, 1]. Estimates outside of this interval serve as an indicator of potentially fraudulently reported votes (Levin et al. 2009).

[22] The focus on the SSD implies that the expected frequencies for digit i are given by $E_i = \{0.120, 0.114, 0.109, 0.104, 0.100, 0.097, 0.090, 0.088, 0.085\}$ and that the chi-square statistic should be compared to the chi-square distribution with nine degrees of freedom.

[23] As Mebane (2008b) notes, it should also be the case that under normal circumstances, the second-digit mean (captured by the intercept) should conform to its *Benford* expected value (4.187).

Table 5 Error rates of fraud detection algorithms in previous research

Procedure	*Correctly classified* (%)	*False positives* (%)	*False negatives* (%)
1BL test	66.6	25.0	0.0
2BL test (Mebane 2008b)	50.0	0.0	50.0
Last-digit test (Beber and Scacco 2008)	50.0	0.0	50.0
Mean of SSD (Mebane 2008b)	25.0	25.0	50.0
Turnout and SSD (Mebane 2008b)	50.0	0.0	50.0
Turnout anomalies (Levin et al. 2009)	50.0	0.0	50.0

Note. This table reports the error rates associated with different fraud detection algorithms used in the literature: the first-digit Benford's Law test (1BL); the second-digit Benford's Law test (2BL) the last-digit test, the SSD's mean test; the second-digit mean/turnout test; and the anomalies in turnout test. An election is considered to be correctly classified if it matches the historical evidence. A false positive (negative) is an election classified as fraudulent (clean) but considered legitimate (irregular) by most historical accounts. Due to data availability constraints, we could only perform the analyses that examine anomalies in the distribution of turnout for the 1940 and 1941 elections.

To carry out the comparison between our *learner* and the six fraud detection algorithms discussed in this section, we used each one of them to examine the electoral contests in our validation/test sets.[24] Recall that because Benford's Law is scale invariant, the Radical party's vote counts would be of little help to establish whether electoral manipulation leads to vote counts that do not satisfy the law. We thus focused on the postmanipulation vote count for the Conservative party. To establish each procedure's classification accuracy, we relied once again on the historical evidence. So, for example, we consider an irregular (legitimate) election to be correctly classified if its deemed fraudulent (clean) by most historical accounts.

Table 5 reports the error rates associated with each of the procedures. In terms of its classification accuracy, the SSD mean test seems to be the weakest. The only election that can be correctly classified using this test is the 1931 provincial contest (as clean). The procedure also produces one false positive (it classifies the 1940 election as fraudulent rather than clean) and two false negatives (the 1935 and 1941 elections, which are classified as clean instead of fraudulent). The remaining tests do not fare much better: They either classify all the analyzed elections as clean (2BL test, Last Digit Test, Turnout and SSD, and Turnout Anomalies) yielding 50% of false negatives or produce at least one false positive (the 1BL test, which classifies the 1931 election as fraudulent).

In contrast to these fraud detection procedures, our *learner* correctly classifies 100% of the cases, with no false positives and no false negatives. As such, the evidence presented in Table 5 suggests that the approach proposed in this paper is a more powerful classification algorithm than the election forensics tools previously used in the literature.

Conclusions

Despite the centrality of elections as mechanisms for providing public accountability, fraud and electoral manipulation remain understudied. Two major limitations have affected the study of electoral fraud. First, there is a dearth in the amount of data that is publicly available to researchers. It is often impossible or at least very difficult to acquire the amount or type of data needed for tests because governments who cheat seldom release fraud figures. The second limitation has been the lack of a widely accepted method to detect electoral fraud when little information is available.

This paper introduces an innovative method to diagnose electoral fraud using recorded vote counts. Building on recent work by Mebane (2006, 2008b) and Beber and Scacco (2008), we rely on digital analysis to identify electoral tampering. We depart from their analyses, however, in a several ways. By doing so, we provide a novel approach for dealing with uncertain data in classification with applications to electoral fraud. First, we develop a method for generating large amounts of synthetic data that preserve statistical properties of a selected set of authentic data used as a seed. Second, we demonstrate that the

[24] Unfortunately, due to data availability constraints, we could only perform the analyses that examine anomalies in the distribution of turnout for the 1940 and 1941 elections.

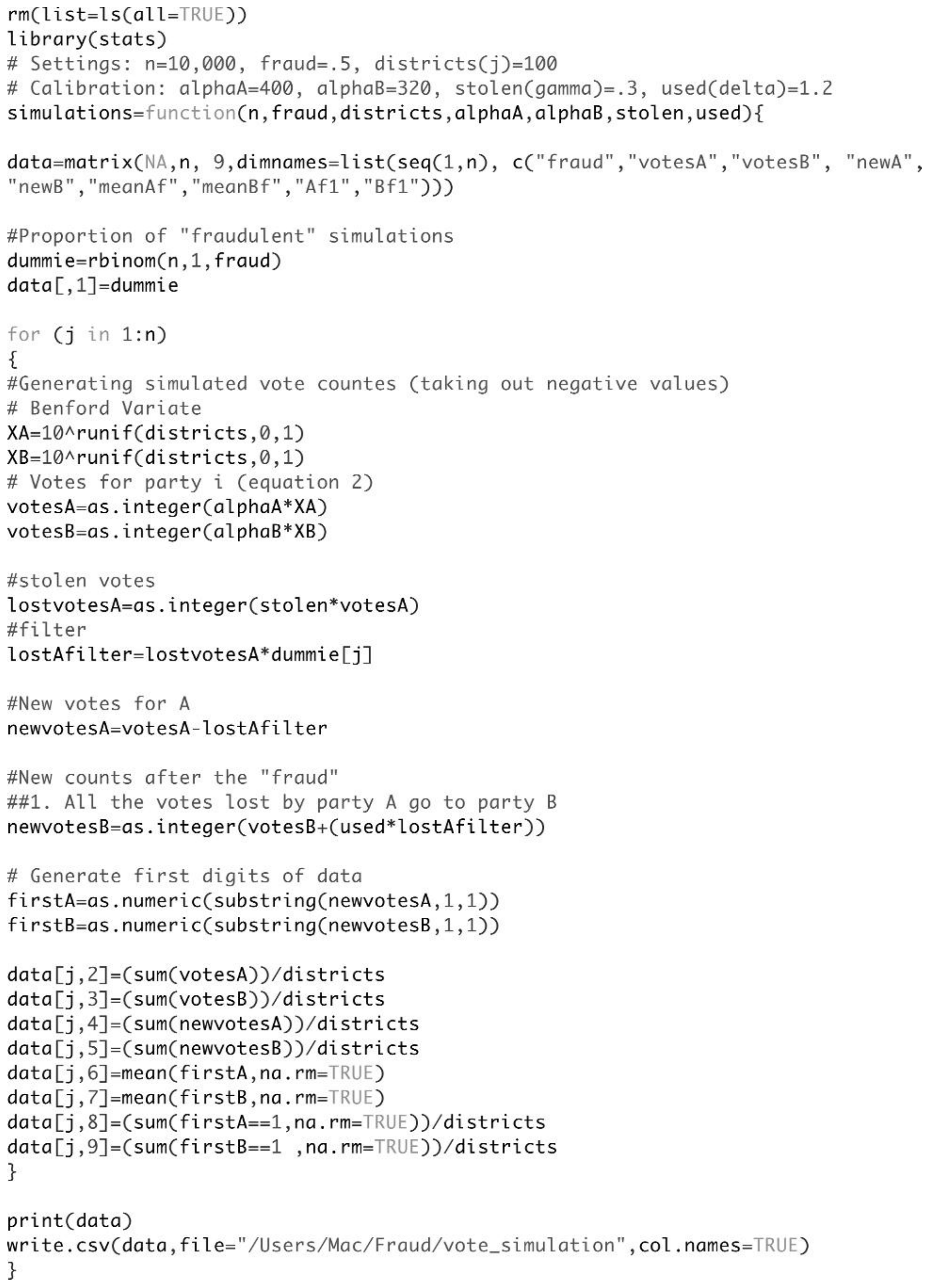
synthetic data can be used to train an electoral fraud detection system. We believe that future studies should consider these methodological innovations when analyzing electoral irregularities.

Substantively, this study provides indisputable evidence of the scope and intensity of electoral fraud during Argentina's "infamous decade." Our findings confirm that electoral fraud, rather than a shift in voters' preferences, led to the dramatic electoral changes during this period. Indeed, our goal was to demonstrate that our approach can be used to distinguish stolen elections from electoral landslides. This concern guided the choices that we made in modeling electoral fraud (the assumption that acts of fraud are committed in every polling station and that they consist of taking votes away from one party to give to the other). Our findings indicate that this assumption accurately depicted the manner in which fraud was committed in Buenos Aires between 1931 and 1941. The appropriate choice, however, ultimately depends on the type of fraud under consideration. Future experiments should verify whether our results also hold for more general classes of seed data and for other types of electoral fraud detection systems.

Data Sources

República Argentina. Congreso Nacional. Cámara de Diputados. *Diario de Sesiones*. Buenos Aires, 1932. Congreso Nacional. Cámara de Diputados. Diario de Sesiones. Buenos Aires, 1936. Ministerio del Interior. *Memoria del ministerio del interior presentada al honorable congreso de la nación*. Buenos Aires, 1935. Ministerio del Interior. *Memoria del ministerio del interior presentada al honorable congreso de la nación*. Buenos Aires, 1941. *Review of the River Plate.* Buenos Aires, 1931–1942.

Appendix 1



```
rm(list=ls(all=TRUE))
library(stats)
# Settings: n=10,000, fraud=.5, districts(j)=100
# Calibration: alphaA=400, alphaB=320, stolen(gamma)=.3, used(delta)=1.2
simulations=function(n,fraud,districts,alphaA,alphaB,stolen,used){

data=matrix(NA,n, 9,dimnames=list(seq(1,n), c("fraud","votesA","votesB", "newA",
"newB","meanAf","meanBf","Af1","Bf1")))

#Proportion of "fraudulent" simulations
dummie=rbinom(n,1,fraud)
data[,1]=dummie

for (j in 1:n)
{
#Generating simulated vote countes (taking out negative values)
# Benford Variate
XA=10^runif(districts,0,1)
XB=10^runif(districts,0,1)
# Votes for party i (equation 2)
votesA=as.integer(alphaA*XA)
votesB=as.integer(alphaB*XB)

#stolen votes
lostvotesA=as.integer(stolen*votesA)
#filter
lostAfilter=lostvotesA*dummie[j]

#New votes for A
newvotesA=votesA-lostAfilter

#New counts after the "fraud"
##1. All the votes lost by party A go to party B
newvotesB=as.integer(votesB+(used*lostAfilter))

# Generate first digits of data
firstA=as.numeric(substring(newvotesA,1,1))
firstB=as.numeric(substring(newvotesB,1,1))

data[j,2]=(sum(votesA))/districts
data[j,3]=(sum(votesB))/districts
data[j,4]=(sum(newvotesA))/districts
data[j,5]=(sum(newvotesB))/districts
data[j,6]=mean(firstA,na.rm=TRUE)
data[j,7]=mean(firstB,na.rm=TRUE)
data[j,8]=(sum(firstA==1,na.rm=TRUE))/districts
data[j,9]=(sum(firstB==1 ,na.rm=TRUE))/districts
}

print(data)
write.csv(data,file="/Users/Mac/Fraud/vote_simulation",col.names=TRUE)
}
```

Appendix 2

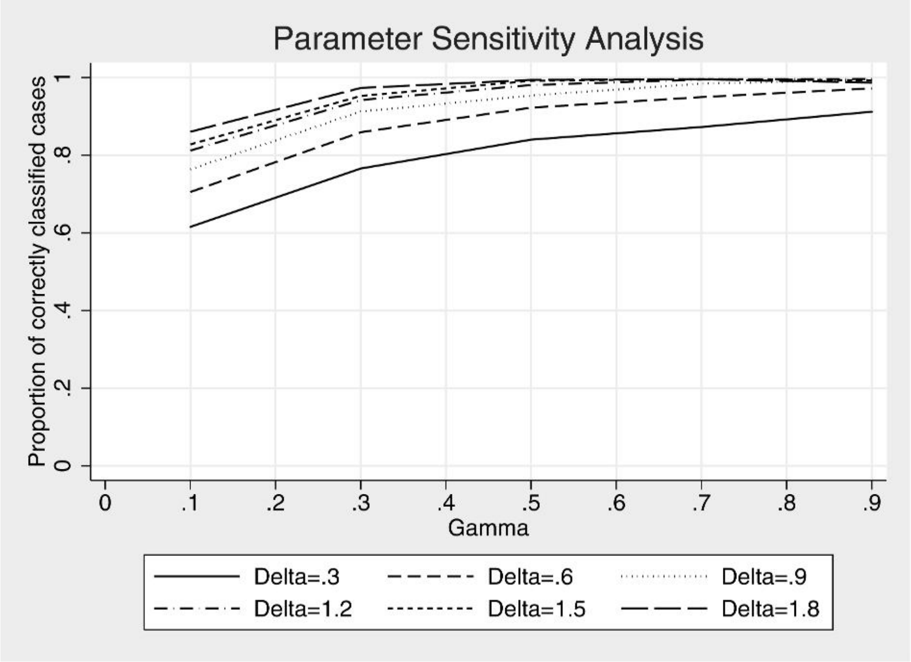


Fig. A2 Sensitivity of classification to simulation assumptions. This figure shows the sensitivity of our *learner*'s classification accuracy to the simulation assumptions. The evidence suggests that even parameter values reflecting different levels of fraud would yield very similar results.

Table A2 Classification of Buenos Aires' elections

Election	$p(clean\|\mathbf{x})$	$p(fraud\|\mathbf{x})$
	$\gamma = 0.1, \delta = 0.6$	
1931	0.533	0.467
1935	0.658	0.342
1940	0.601	0.399
1941	0.640	0.360
	Correctly classified: 50%	
	$\gamma = 0.5, \delta = 1.8$	
1931	0.475	0.525
1935	0.045	0.955
1940	0.300	0.700
1941	0.066	0.934
	Correctly classified: 50%	

Note. This table confirms that appropriateness of our simulation parameters ($\gamma = 0.3$, and $\delta = 1.2$). When the simulated amount of fraud is scant ($\gamma = 0.1$, $\delta = 0.6$), the classifier's performance regarding the Buenos Aires elections suffers significantly: All elections are considered clean. Likewise, if we simulate too much fraud ($\gamma = 0.5$, $\delta = 1.8$), the classifier also performs quite badly, but in the opposite direction: All elections are considered fraudulent.

Appendix 3

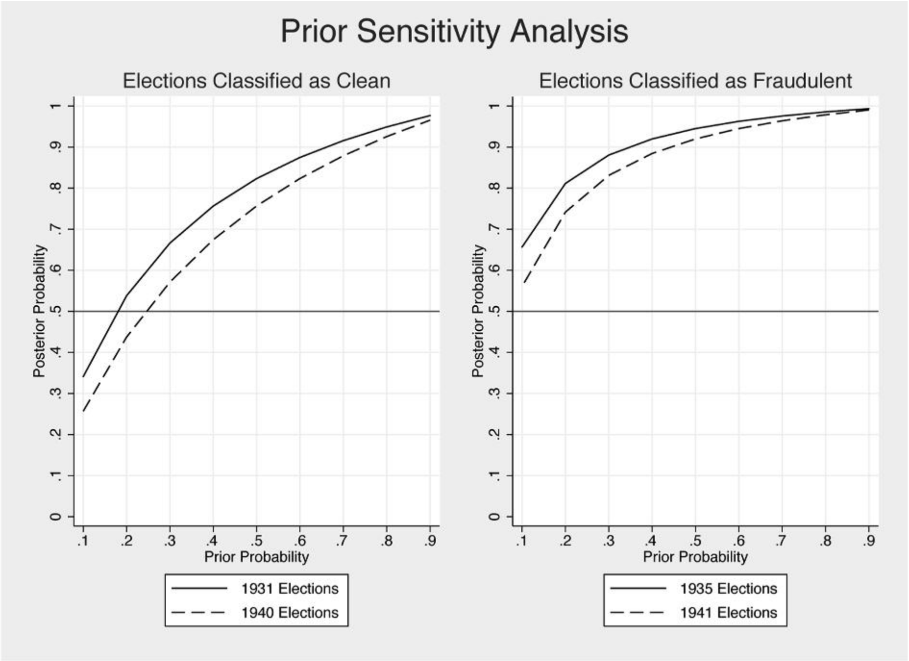


Fig. A3 Sensitivity of classification accuracy to priors. This figure shows the sensitivity of our *learner*'s classification accuracy to different assumptions regarding the prior probabilities of observing a fraudulent/clean election. The findings indicate that our result are robust to different expectations. In the case of fraudulent elections, correctly classified outcomes entail posterior probabilities higher than .5, for the whole range of priors and for clean elections for prior probabilities, $p(\text{clean}) > .25$.

References

Abal Medina, Juan Manuel, and Julieta Suárez Cao. 2003. Partisan competition in Argentina. From closed and predictable to open and unpredictable. Meeting of the Latin American Studies Association, Dallas, TX.

Alston, Lee J., and Andres A. Gallo. 2010. Electoral fraud, the rise of Peron and demise of checks and balances in Argentina. *Explorations in Economic History* 47:179–97.

Altinçay, Hakan. 2005. On naive Bayesian fusion of dependent classifiers. *Pattern Recognition Letters* 26:2463–73.

Alvarez, R. Michael, Thad E. Hall, and Susan D. Hyde. 2008. *Election fraud: Detecting and deterring electoral manipulation*. New York: The Brookings Institution.

Beber, Bernd, and Alexandra Scacco. 2008. What the numbers say: A digit-based test for election fraud using new data from Nigeria. Working Paper.

Bejar, María Dolores. 2005. *El Régimen Fraudulento. La política en la provincia de Buenos Aires, 1930–1943*. Buenos Aires: Siglo XXI editores.

Busta, Bruce, and Randy Weinberg. 1998. Using Benford's Law and neural networks as review procedure. *Managerial Auditing Journal* 13:356–66.

Bustamante, Carlos, Leonardo Garrido, and Rogelio Soto. 2006. Comparing fuzzy naive Bayes and Gaussian naive Bayes for decision making in RoboCup 3D. In *Advances in artificial intelligence*. Berlin, Germany: Springer.

Buttorf, Gail. 2008. Detecting fraud in America's gilded age. Working Paper.

Chan, P., W. Fan, A. Prodromiris, and S. Stolfo. 1999. Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems* 14:67–74.

Cho, Wendy K. Tam, and Brian J. Gaines. 2007. Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American Statistician* 61:218–23.

Ciofalo, Michele. 2009. *Entropy, Benford's first digit law, and the distribution of everything*. Palermo, Italy: Dipartamento di Ingenieria Nucleare, Universita degli Studi di Palermo.

Ciria, Alberto. 1974. *Parties and power in modern Argentina*. Albany, NY: State University of New York Press.

Clifford, P., and A. F. Heath. 1993. The political consequences of social mobility. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 156:51–61.

Cox, Gary W. 1997. *Making votes count*. New York: Cambridge University Press.

Cox, Gary W., and Morgan Kousser. 1981. Turnout and rural corruption: New York as a test case. *American Journal of Political Science* 25:646–63.

Debar, H., M. Dacier, A. Wespi, and S. Lampart. 1998. An experimentation workbench for intrusion detection systems. Technical Report RZ2998. Zurich, Switzerland: IBM Research Division.

Demichelis, Francesca, Paolo Magni, Paolo Piergiorgi, Mark Rubin, and Riccardo Bellazzi. 2006. A hierarchical Naive Bayes Model for handling sample heterogeneity in classification problems: An application to tissue microarrays. *BMC Bioinformatics* 514(7):1–12.

Demirekler, Mübeccel, and Hakan Altinçay. 2002. Plurality voting-based multiple classifier systems: Statistically independent with respect to dependent classifier sets. *Pattern Recognition* 35:2365–79.

Domingos, Pedro, and Michael Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29(2–3):103–30.

Drake, Paul. 2009. *Between tyranny and anarchy*. Stanford, CA: Stanford University Press.

Drake, Philip D., and Mark J. Nigrini. 2000. Computer assisted analytical procedures using Benford's Law. *Journal of Accounting Education* 18:127–46.

Eno, Josh, and Craig W. Thompson. 2008. Generating synthetic data to match mininig patterns. *Internet Computing* 12:78–82.

Fewster, R. M. 2009. A simple explanation of Benford's Law. *The American Statistician* 63:26–32.

Grendar, Marian, George Judge, and Laura Schechter. 2007. An empirical non-parametric likelihood family of data-based Benford-like distributions. *Physica A* 380:429–38.

Haines, J. W., R. P. Lippmann, D. J. Fried, E. Tran, S. Boswell, and M. A. Zissman. 2001. Data intrusion detection system evaluation: Design and procedures. Technical Report 1062. Lexington, MA: MIT Lincoln Laboratory.

Hand, David J., and Keming Yu. 2001. Idiot's Bayes—Not so stupid after all? *International Statistical Review* 3:385–98.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning*. New York: Springer.

Hill, Theodore P. 1995a. The significant-digit phenomenon. *The American Mathematical Monthly* 102:322–7.

———. 1995b. A statistical derivation of the significant-digit law. *Statistical Science* 10:354–63.

Hyde, Susan D. 2007. The observer effect in international politics: Evidence from a natural experiment. *World Politics* 50(1): 37–63.

Janvresse, Élise, and Thierry de la Rue. 2004. From uniform distributions to Benford's Law. *Journal of Applied Probability* 41: 1203–10.

Katz, Jonathan N., and Brian R. Sala. 1996. Careerism, committee assignments, and the electoral connection. *The American Political Science Review* 90:21–33.

Kotsiantis, S. B. 2007. Supervised machine learning: A review of classification techniques. *Informatica* 31:249–68.

Kuncheva, Ludmila I. 2006. On the optimality of Naive Bayes with dependent binary features. *Pattern Recognition Letters* 27: 830–7.

Kvarnstrom, H., E. Lundin, and E. Jonsson. 2000. Combining fraud and intrusion detection—Meeting new requirements. Proceedings of the Fifth Nordic Workshop on Secure IT systems, Reykjavik, Iceland, October 12–13.

Leemis, Lawrence M., Bruce Schmeiser, and Diane L. Evans. 2000. Survival distributions satisfying Benford's Law. *The American Statistician* 54:236–41.

Lehoucq, Fabrice. 2003. Electoral fraud: Causes, types, and consequences. *Annual Review of Political Science* 6:233–56.

Levin, Ines, Gabe Cohn, R. Michael Alvarez, and Peter C. Ordeshook. 2009. Detecting voter fraud in an electronic voting context: An analysis of the unlimited reelection vote in Venezuela. Online Proceedings of the Electronic Voting Technology Workshop.

Lundin, Emilie, Håkan Kvarnström, and Erland Jonsson. 2002. *A Synthetic Fraud Data Generation Methodology*. Lecture Notes in Computer Science. Berlin, Germany: Springer.

Lupu, Noam, and Susan Stokes. 2009. The social bases of political parties in Argentina, 1912–2003. *Latin American Research Review* 44:58–87.

Mebane, Walter R. 2006. Election forensics: Vote counts and Benford's Law. 2006 Summer Meeting of the Political Methodology Society, UC-Davis.

———. 2007. Statistics for digits. 2007 Summer Meeting of the Political Methodology Society, Penn State University, University Park, PA.

———. 2008a. Election forensics: Outlier and digit tests in America and Russia. Working Paper.

———. 2008b. Elections forensics: The second-digit Benford Law's test and recent American presidential elections. In *Election fraud*, ed. R. Michael Alvarez, Thad E. Hall, and Susan D. Hyde. Washington, DC: The Brookings Institutions.

———. 2010. Fraud in the 2009 presidential election in Iran? *Chance* 23:6–15.

Mitchell, T. 1997. *Machine learning*. New York: McGraw-Hill.

Nye, John, and Charles Moul. 2007. The political economy of numbers: On the application of Benford's Law to international macroeconomics statistics. *The B.E. Journal of Macroeconomics* 7(1):1–12.

Pericchi, Luis R., and David Torres. 2004. La Ley de Newcomb-Benford y sus aplicaciones al Referendum Revocatorio en Venezuela. Working Paper.

Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler. 2005. A comprehensive survey of data mining-based fraud detection research. *Victoria*:1–14. http://arxiv.org/abs/1009.6119.

Potash, Robert A. 1969. *The army and politics in Argentina: 1928–1945*. Stanford, CA: Stanford University Press.

Przeworski, Adam. 2010. *Democracy and the limits of self-government*. New York: Cambridge University Press.

Puketza, N. J., K. Zhang, M. Chung, B. Mukherjee, and R. A. Olsson. 1996. A methodology for testing intrusion detection systems. *Software Engineering* 22(10):719–29.

Reiter, J. P. 2004. Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* 30: 235–42.

Rish, Irina. 2001. An empirical study of the naive Bayes classifier. Proceedings of the IJCAI workshop on "Empirical Methods in AI."

Roukema, Boudewijn F. 2009. Benford's Law anomalies in the 2009 Iranian presidential election. Unpublished manuscript.

Rubin, Donald B. 1993. Discussion statistical disclosure limitation. *Journal of Official Statistics* 9:461–8.

Schäfer, Christin, Jörg-Peter Schräpler, Klaus-Robert Müller, and Gert G. Wagner. 2004. Automatic identification of faked and fraudulent interviews in surveys by two different methods. Working Paper.

Tan, Aik Choon, and David Gilbert. 2003. An empirical comparison of supervised machine learning techniques in bioinformatics. First Asia-Pacific Bioinformatics Conference, Adelaide, Australia, February 4–7.

Varian, Hal A. 1972. Benford's Law. *The American Statistician* 26(3):65–6.

Walter, Richard J. 1985. *The province of Buenos Aires and Argentine politics, 1912–1943*. New York: Cambridge University Press.

Wong, Weng-Keen, Andrew Moore, Gregory Cooper, and Michael Wagner. 2003. Bayesian network anomaly pattern detection for disease outbreaks. Proceedings of the International Conference on Machine Learning, Washington, DC, August 21–24.

Yu, Lei, and Huan Liu. 2004. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* 5:1205–24.

Zetter, Kim. 2009. Crunching Iranian election numbers for evidence of fraud. http://www.wired.com/threatlevel/2009/06/iran_numbers.

Zhang, Harry. 2004. The optimality of naive Bayes. Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference. The AAAI Press, pp. 562–67.

Zhang, Harry, and Jiang Su. 2004. Naive Bayesian classifiers for ranking. In *ECML 2004, lecture notes in Computer Science*, ed. J.-F. Boulicaut. Berlin, Germany: Springer.