Election Fraud: A Latent Class Framework for Digit-Based Tests

Juraj Medzihorsky

Thursday 30th July, 2015

ABSTRACT

Digit-based election forensics typically relies on null hypothesis significance testing, with undesirable effects on substantive conclusions. This paper proposes an alternative free of this problem. It rests on decomposing the observed numeral distribution into the 'no fraud' and 'fraud' latent classes, by finding the smallest fraction of numerals that either needs to be removed or reallocated to achieve a perfect fit of the 'no fraud' model. The size of this fraction can be interpreted as a measure of fraudulence. Both alternatives are special cases of measures of model fit–the π^* mixture index of fit and the Δ dissimilarity index, respectively. Furthermore, independently of the latent class framework, the distributional assumptions of digit-based election forensics can be relaxed in some contexts. Independently or jointly, the latent class framework and the relaxed distributional assumptions allow to dissect the observed distributions using models more flexible than those of existing digit-based election forensics. Reanalysis of Beber and Scacco's (2012) data shows that the approach can lead to new substantive conclusions.

Author's note: I am grateful to Tamás Rudas, Gábor Tóka, Levente Littvay, Zoltán Fazekas, Daniela Širinić, Pavol Hardos, two anonymous reviewers, and the editors for helpful comments and suggestions, and the members of the Political Behavior Research Group at CEU for a helpful discussion. Replication materials are available online as Medzihorsky, Juraj, 2015, "Replication Data for: Election Fraud: A Latent Class Framework for Digit-Based Tests", `http://dx.doi.org/10.7910/DVN/1FYXUJ`, Harvard Dataverse, V1 [UNF:6:FIWHvsHNzZgPStT0+kgbsQ==] (Medzihorsky, 2015b), and include the version of the `R` package `pistar` (Medzihorsky, 2015a) used in the analysis. The article uses data from Beber and Scacco (2012) which is available online also as (Beber and Scacco, 2011). Supplementary materials for this article are available on the *Political Analysis* Web site.

1. INTRODUCTION

Methods for electoral integrity evaluation include statistical ones, some of which focus on digit distributions in electoral returns and are known as 'election forensics.'[1] Digit-based election forensics (DBEF) promises to detect some kinds of electoral fraud by an inexpensive inspection of the electoral returns. Two features typify it. The first is the assumption that the distribution of numerals in fraud-free returns is known and different from that in fraudulent ones, hereafter referred to as the *strong distributional assumption.* The second is the evaluation of elections by testing the statistical significance of the deviation from the null hypothesis that the observed results are fraud-free.

The validity of DBEF has been questioned, mainly due to its distributional assumption. Independently of this, there are issues with the use of null hypothesis significance testing (NHST) detrimental to the usefulness of DBEF. The main contribution of the present paper is an alternative statistical approach based on latent classes free of these issues. The secondary contribution is an independent method to relax the strong distributional assumption in some contexts. The main finding is that adopting one or both can lead to new substantive conclusions, and provide a new perspective on the enterprise of DBEF.

The core assumptions of DBEF and the criticism leveled at them are summarized in Section 2. Section 3 critiques the use of NHST in DBEF, and Section 4 introduces a latent class based alternative. An independent approach to relax the distributional assumption of DBEF implemented with loglinear models is presented in Section 5. Section 6 shows, reanalyzing Beber and Scacco's (2012) data, that the proposed methods can lead to new substantive conclusions as well as a new perspective on the enterprise DBEF.

[1] The term 'election forensics' has been coined by Mebane (2006a). Other examples of this approach are found e.g. in Mebane (2006b, 2007, 2008, 2010a,b); Mebane and Kalinin (2009); Buttorf (2008); Breunig and Goerres (2011); Pericchi and Torres (2011); Beber and Scacco (2012), and as a part of a more general approach in **?**. For an overview of electoral integrity evaluation methods and the place of election forensics in this context see **?** and Alvarez et al. (2012).

2. DISTRIBUTIONAL ASSUMPTIONS IN DBEF

Existing DBEF methods are based on the assumption that the distribution of numerals in fraud-free vote counts is known and different from that in fraudulent ones. Typically, this distribution is derived from Benford's law (BL), an observation that for certain kinds numbers the frequencies of digits at each position resemble a logarithmic distribution (Newcomb, 1881; Benford, 1938). Under BL the probability of leading digit $d \in \{1, \dots, 9\}$ is

$$P(D_1 = d) = \log_{10}\left(1 + \frac{1}{d}\right),$$

and of digit $d \in \{0, \dots, 9\}$ in position $j \in \{2, \dots, J\}$

$$P(D_j = d) = \sum_{k=10^{j-2}}^{10^{j-1}-1} \log_{10}\left(1 + \frac{1}{10k + d}\right).$$

Not all DBEF uses BL–Beber and Scacco (2012) expect under no fraud uniformly distributed numerals in the last digits of three or more digit numbers. In practice, this is not radically different, since the Benford distribution with increasing digit order approaches uniformity, getting close to it already at the third digit (see Table A.1 in the Appendix).

The strong distributional assumption of DBEF has been subject to several critiques. Since the present paper proposes a way to bypass these issues in some contexts, they are only briefly summarized here. The relevance of BL for fraud-free vote counts is argued for using formal and empirical evidence. First, it has been proved that digits in numbers randomly drawn from a random mixture of distributions converge to a logarithmic distribution (Hill, 1995), and observed that many numbers that represent naturally occurring phenomena follow it. Second, in applications such as forensic accounting and scientific fraud detection, BL-based methods have been considered successful.[2] It is contentious that elections, free of fraud or not, share the rele-

[2]A brief discussion of digit-based forensics outside of elections is provided by Beber and Scacco (2012). Inter-

vant features with such processes (Deckert et al., 2011). Third, studies that attempt to simulate fraud-free elections show that numerals in the simulated vote counts follow BL (e.g. Mebane, 2006a) or uniformity (Beber and Scacco, 2012). However, there is little evidence that the simulated processes bear sufficient resemblance to their empirical counterparts, and given the lack of well-formed models the value of these simulations is not known (see Deckert et al. 2011 and also Mebane 2011). Fourth, some elections for which there is considerable other evidence of fraud or its absence are evaluated congruently by DBEF (Beber and Scacco, 2012). The value of this validation depends on whether the sample represents typical fraudulent and fraud-free elections, which at best has a large uncertainty attached.

The expectation that BL (or uniformity) will not hold under fraud is argued for using evidence from experiments and simulations. First, a considerable amount of experimental evidence shows that when asked to generate random numbers, humans produce numbers with numeral distributions different from Benford's or uniform (see e.g. Nickerson, 2002; Beber and Scacco, 2012). Since fabrication of electoral returns is a similar process, it is argued, it should show in the distribution of the digits. The experiments have used a range of subjects of different nationalities, ages, and education. Yet, it is not known how well do their findings extend to those with very different backgrounds, or to fraudsters who operate under different priorities and constraints. Also, there is little to prevent the fraudsters from using simple tools such as dice or pseudo-random number generators. Second, simulated fraudulent elections produce vote counts with numerals that do not follow BL (e.g. Mebane, 2006a). It merits skepticism whether the simulations adequately represent their empirical counterparts.

These issues are crucial for the validity of the existing DBEF methods, and Section 5 shows how they can be in some contexts bypassed. Yet, there are also issues related to the statistical techniques used in DBEF that are of equal importance. These are introduced in the following section, and resolved in Section 4.

---

ested readers can find a comprehensive inter-disciplinary bibliography on Benford's law and its application at the Benford Online Bibliography `http://benfordonline.net/list/alphabetical`.

3. STATISTICAL ISSUES IN DBEF

Existing DBEF studies differ in the order of the inspected digits, their expected distributions, and the statistics used. Yet, all compare the observed digit distribution to the one expected under no fraud, typically by testing the statistical significance of the deviation from the null hypothesis that the observations are drawn from the expected distribution. The null hypothesis significance testing framework is a venerable statistical workhorse, fruitfully applied to a wide range of problems. However, a large and diverse body of literature finds issues with its features, often repeatedly and independently (see e.g. Ziliak and McCloskey, 2008). Several of these features have within the context of DBEF unappealing effects on substantive inferences.

First, NHST assumes that the data were generated by stochastic sampling. However, elections are unique events that cannot even in principle be repeated, and their returns are better understood as population data. Relatedly, it is embedded in NHST that with a known frequency the test will reject the null hypothesis when it is true (i.e., commit a Type I error). In the context of DBEF it means that some non-fraudulent elections will be labeled as fraudulent, and regardless of whether fraud is present, the more the investigator tries to uncover it, the more evidence of it she will find. Since allegations of fraud often arise in new democracies and developing countries (see e.g. Norris et al., 2014), and can reflect the biases of actors that make them, this can lead to the confirmation of these biases. Lowering the Type I error rate comes at the price of the power of the test, and its choice should be based on operational concerns. Since fair elections are fundamental to democratic legitimacy, and raising and amplifying unsubstantiated suspicions of electoral fraud can undermine it, setting the rate is at best an extremely sensitive task. In short, NHST is designed for settings where stochastic samples are repeatedly taken from a population, and the Type I error rate is chosen based on operational concerns, neither of which applies well in DBEF.

Third, inferences are limited if the test rejects the model. The test is not informative if the model is not true, because it rests on the comparison of the value of the test statistic with its distribution if the model is true. Furthermore, if all restrictive models are rejected, there are no validly defined residuals. In existing DBEF the restrictive models are of fraud-free elections, and no alternative ones are considered. Thus, if the model of no fraud is rejected, there is no alternative model, and no residuals.

Finally, test statistics have features that can be unappealing in some contexts. Perhaps the best known test statistic, Pearson χ^2, is sensitive to sample size in the sense that it might lead to different conclusions for two samples with identical densities, but different sample sizes. In the context of DBEF this might lead to rejections of models of no fraud with large samples. This has been recognized in the digit based forensic literature and different solutions have been devised.

One group of solutions rests on offering different test statistics (Leemis et al., 2000; Giles, 2007; Tam Cho and Gaines, 2007; Judge and Schechter, 2009). As illustrated by Judge and Schechter's (2009) analysis, these statistics can with the same data on the same level of statistical significance lead to different conclusions. Simply, alternative test statistics offer different trade-offs in terms of their sensitivity, but do not resolve the above mentioned issues embedded in the NHST framework.

Another group of solutions relies on Bayesian inference. Pericchi and Torres (2011) and Jiménez and Hidalgo (2014) build on the χ^2 test, but use adjusted p-values and compute Bayes factors, obtaining an appealing quantity–the probability of the hypothesis given the data. This approach builds on the χ^2 test and inherits some of its assumptions, and furthermore requires the use of priors. Cantú and Saiegh (2011) use a naive Bayes classifier, trained on synthetic data using significance testing, and generate for each inspected set of digits a posterior probability of being fraudulent. In contrast, the latent class approach introduced in the present paper does not require the use of priors, does not rely on NHST in any way, and rests on assumptions known to be true.

The above outlined issues stem from the fundamental setup of the null hypothesis significance testing framework, and cannot be resolved by choices within it, such as of a different test statistic or significance level. The next section outlines an alternative framework based on latent class analysis that can replace NHST in DBEF.

4. LATENT CLASS DIGIT BASED ELECTION FORENSICS

In a set of electoral returns fraud can affect between 0% and 100% of the reported numbers. Accordingly, the observed distribution of numerals O is composed of a distribution N that belongs to the non-fraudulent class, and a distribution F that belongs to the fraudulent one, the size of which $\zeta \in [0,1]$,

$$O = (1-\zeta)N + \zeta F. \tag{1}$$

Existing DBEF methods restrict the digit distribution only for the non-fraudulent results, leaving that of the fraudulent ones unspecified. Because of this flexibility, the model in (1) describes the data perfectly under values of ζ on $[\zeta_L, 1]$, where ζ_L is the lowest such size of the unrestricted component that will still result in perfect fit. If a non-fraudulent distribution describes the data perfectly, then ζ_L is zero. The value of ζ_L depends on whether to achieve a perfect fit of the no-fraud model the unrestricted component is to be removed or reallocated. In the former case ζ_L is a special case of the π^* mixture index of fit (Rudas et al., 1994) and in the latter of the Δ dissimilarity index (Gini, 1914). Standard measures of model fit assume that a single model describes the whole population. The indexes abandon this idea, and assume that the population is composed of two classes, units for which the model holds true and those for which it does not.

4.1. *The π^* Mixture Index of Fit*

If the perfect fit is to be achieved by removing the unrestricted component, then under the scaled distribution of the non-fraudulent class $(1-\zeta)N$ the probability of any of the digits is not higher than its observed one. After substituting ζ with π, for any set of observed digits the fit of this model is perfect for all π on $[\pi^*, 1]$, where π^* is the lowest such size that will still result in perfect fit. This quantity can be interpreted as the smallest fraction of the inspected digits that cannot be described as free of fraud. As such, it is a special case of the π^* mixture index of fit (Rudas et al., 1994; Rudas, 1998, 1999, 2002), a latent class based measure of model fit with wide applicability.[3]

The quantity of interest is π^*, the smallest share of cases for which the model does not hold. It can be understood as a measure of distance from the observations O to the model $\mathscr{M}$, the smallest such π that decomposes the observed density perfectly into an element M from the model and an unspecified component U,

$$\pi^*(O, \mathscr{M}) = \inf\{\pi\colon\ O = (1-\pi)M + \pi U,\ M \in \mathscr{M},\ U \text{ unspecified}\}.$$

Conventional models of fraud-free digit distributions lack free parameters, which eases the application of the index. To obtain the value of the index, the scaled model density $(1-\pi)M$ needs to be 'below' the observed density O while 'shrinking' as little as possible, that is only as much as it needs to fit in the cell where it fits the worst. This is the cell with the highest ratio of the model density over the observed density. The solution is to multiply the model density by the

[3]The π^* index of fit has been applied in a variety of settings including contingency tables analysis (Rudas et al., 1994; Clogg et al., 1995, 1997), item response models (Rudas and Zwick, 1997; Hernández et al., 2006; Formann, 2006; Revuelta, 2008), latent class analysis (Formann, 2000, 2003a,b), regression (Rudas, 1999; Verdes and Rudas, 2003), missing data (Rudas, 2005; Rudas and Verdes, 2015), and robust statistics (Ispány and Verdes, 2014).

inverse of this ratio. The value of π^* is

$$\pi^* = 1 - \frac{1}{\max\limits_{i=1,\dots,N} \left\{ \frac{M_i}{O_i} \right\}},$$

as shown more generally by Rudas (1999). The procedure is illustrated in Figure 1 with simulated data.

[Figure 1 about here.]

4.2. *The Δ Dissimilarity Index*

If the perfect fit of the model is to be achieved by reallocating some of the observations, then ζ_L in (1) is the sum of the absolute values of the residuals divided by twice the sample size. This is because the residuals are symmetric in the sense that the sum of the absolute values of the positive residuals is equal to that of the negative residuals. In this setting ζ_L is a special case of the Δ dissimilarity index (Gini, 1914), defined for a contingency table with cell counts O_i, predicted values M_i, and sample size N as

$$\Delta = \frac{\sum_{i=1}^{N} |O_i - M_i|}{2N}.$$

While the Δ index is not usually presented in latent class terms, in that case the interpretation of the in-model component is different than under the π^* index. Whereas under π^* the unscaled 'in-model' distribution N is perfectly described by the model, it is not under Δ. Instead, the scaled 'in-model' distribution $(1-\zeta)N$ is interpreted as the largest such fraction that does not need to be reallocated to achieve a perfect fit of the model if the rest (ζF) is reallocated to this aim.

The dissimilarity index can be given a straightforward interpretation in the context of electoral fraud. In an election, usually the returns for each of the options from each territory have to be reported. Thus, also the total number of their last digits is fixed. If the fraudsters substitute a true number by a fraudulent one, its last digit can be changed and thus reallocated. Then, Δ can be interpreted as the smallest fraction of digits that would need to be changed to their presumed original values in order to observe the distribution thought to characterize the absence of fraud.

4.3. *The Appeal of π^* and Δ in Election Forensics*

In the context of digit-based election forensics the π^* and Δ indexes are appealing with a host of features.[4] First, they do not assume homogeneity or stochastic sampling, and consequently their underlying assumptions are always true. Electoral returns are better understood as a population, and this allows to treat them as such. Second, they do not rest on rejecting or retaining a null hypothesis, and do not run the risks of Type I and II errors. Consequently, regardless of how many elections will be inspected, none will be falsely labeled as fraudulent or fraud-free. The amount of fraud will simply be over- or under-estimated for some. In short, the indexes allow to ask the quantitative question of how much fraud there was, instead of just the qualitative one of whether there was fraud.

Third, few model fit statistics have an equally straightforward interpretation and are easy to reason about as do π^* and Δ. Fourth, both indexes are independent of sample size in the sense that for any of them the value is the same for any two datasets with the same observed probability distribution regardless of their size. This allows to compare the fit of the model to datasets of different sizes. Fifth, the units in the unrestricted component can be interpreted as residuals. Thus, unlike in the conventional approach, the residuals are defined in a way that is always valid, and are available for interpretation regardless of how badly does the model fit (see esp. Clogg et al., 1995).

[4]These features are discussed in detail in the context of π^* by Rudas et al. (1994); Clogg et al. (1995), and Rudas (1998, 2002), with the exception of the use of jackknife and the use as a test statistic in NHST.

Additional inferential leverage is available under stochastic sampling. The uncertainty attached to sample size can be represented with confidence intervals for the fit statistic, e.g. via jackknife, as shown for π^* by Dayton (2003). From a strict frequentist perspective, if the data are not a stochastic sample, such interval estimates are not meaningful. From a pragmatically Bayesian perspective, given their low cost they can be treated as approximations of credible intervals under weakly informative priors. In addition, the indexes can serve as test statistics in NHST using simulated reference distributions. Given the discussed features of NHST this might not be a procedure of first choice.

Finally, the present paper applies the indexes by decomposing the observations into a component that belongs to a restrictive model and an unrestricted component. The decrease in power due to the use of the unrestricted component is offset by considerably lighter assumptions. Should restrictive models of fraud processes be available, the indexes can be applied either to them, or to mixtures of models of fraudulent and fraud-free processes. In such settings the indexes could no longer be automatically interpreted as degrees of fraudulence, but the other advantages they have over NHST would remain.

5. RELAXING THE DISTRIBUTIONAL ASSUMPTIONS OF DBEF

Most existing criticism of DBEF is directed at its strong distributional assumption. This assumption can be relaxed in a way independent of the proposed latent class approach. Specifically, where multiple sets of numerals are available, the investigator can ask whether they can be described by the same probability distribution under the *relaxed distributional assumption* that numerals in fraudulent and fraud-free results are distributed differently. Under this assumption, if a single probability distribution describes all the relevant subsets, either all or none of them are deemed fraudulent. Other evidence can decide which of these interpretations is more appropriate. The ability of such procedures to detect fraud depends on more evidence than the existing

DBEF, but with much weaker and less controversial assumptions.

[Table 1 about here.]

Statistically, such inspections can be done in multiple ways, including loglinear models (see e.g. Agresti, 2002, 314-56), which are appealing due to their flexibility. An intuition can be gained from the following example. In an election suspect of fraud observers were deployed, and reported little evidence of fraud in the visited wards. This data can be represented as a contingency table shown in Table 1, where d_{ij} is the count if i^{th} numeral in the j^{th} category. The simplest loglinear model which can applied is

$$\log d_{ij} = \lambda,$$

under which all numerals are equally likely regardless of whether they are from a ward with observers. If the proportion of wards with observers is not 50%, it is useful to introduce the 'observer' parameters λ_j,

$$\log d_{ij} = \lambda + \lambda_j,$$

which is equivalent to expecting the digits to be uniformly distributed within each group of wards, but not necessarily overall. The model of independence for this data is

$$\log d_{ij} = \lambda + \lambda_i + \lambda_j,$$

which allows also the frequency of each numeral to differ by including the numeral parameters λ_i. Under this model a single probability distribution estimated from the data describes the numerals, regardless of whether they came from the wards with observers. Assuming that fraud manifests itself in digit distributions, and that the observer reports are highly credible, this would temper suspicions of fraud. If the model does not fit well, under the same assumptions the allegations of fraud would appear more credible.

Typically, the situation tends to be more complex than in the above example. Most electoral returns contain at least the following information on the digits of interest–numeral, party (candidate), and who carried a given territory. This allows to use the independence model

$$\log d_{ijk} = \lambda + \lambda_i^D + \lambda_j^P + \lambda_k^R,$$

where λ_i^D are the numeral (digit), λ_j^P the party, and λ_k^R the result parameters. The usefulness of this model is limited as it restricts the numbers of won/lost territories to be the same for all parties. Substantively interesting models will lie between the independence model and the saturated model

$$\log d_{ijk} = \lambda + \lambda_i^D + \lambda_j^P + \lambda_k^R + \lambda_{ij}^{DP} + \lambda_{ik}^{DR} + \lambda_{jk}^{PR} + \lambda_{ijk}^{DPR},$$

which fits perfectly by definition. Substantive inferences can be drawn by comparing the fit of the models that include one or more of the two-factor interaction (association) terms $\{\lambda_{ij}^{DP}, \lambda_{ik}^{DR}, \lambda_{jk}^{PR}\}$. If a model without association terms involving numerals describes the data well, then a single probability distribution fits the numerals under all party-result combinations. The use of this approach is shown in the following section on empirical examples of elections believed to be fraudulent as well as fraud-free.

The above examples deal in accord with the focus of this paper with strictly descriptive questions of differences between distributions, and by extension of presence of fraud. However, in some settings the loglinear analysis can be extended to handle causal questions. For instance, considering the first example, conditional on the observer-allocation procedure and other available data, it might be possible to estimate the effect of observer presence on electoral integrity. Loglinear models can be used to test hypothesized causes of fraud, provided the variables of interest are either discrete, or can be discretized with acceptable loss of information.

6. EMPIRICAL DEMONSTRATION

Beber and Scacco (2012) validate their diagnostic procedures using both elections strongly suspect of fraud and elections widely considered fraud-free.[5] This section reanalyzes their data with the proposed approach as well as with a conventional approach, represented by Pearson's χ^2 test, the most common NHST method in the digit-based forensic literature. I opt for the no-fraud model of uniformity, since as Beber and Scacco (2012) demonstrate, it is better theoretically founded than the Benford-based alternatives. Each set of electoral returns is first analyzed under the strong distributional assumption of DBEF and then relaxing this assumption.

6.1. *Sweden 2002*

The Swedish parliamentary elections of 2002 were selected by Beber and Scacco (2012) as an example of elections believed to be fraud-free, and identified as such by them. In the reanalysis, I classify the last digits by numeral, party, and ward result. Detailed results are reported in the Appendix, Section A.3.1. Under the strong distributional assumption both latent class indexes show that uniformity describes the inspected sets of numerals well and thus no suspicions of fraud are raised. NHST leads to the same conclusion, but at the price of more difficult assumptions. Under the relaxed distributional assumption the extent of contamination by fraud seems similar in the inspected subsets. Given other evidence that indicates the absence of fraud, we can conclude that this extent is practically zero, and the detected small departures from the model are due to other causes.

[5] Beber and Scacco's (2012) data are available online as Beber and Scacco (2011). The data used in the analysis, as well as the replication code are available online Replication materials are available online as Medzihorsky, Juraj, 2015, "Replication Data for: Election Fraud: A Latent Class Framework for Digit-Based Tests", `http://dx.doi.org/10.7910/DVN/1FYXUJ`, Harvard Dataverse, V1 [UNF:6:FIWHvsHNzZgPStT0+kgbsQ==] (Medzihorsky, 2015b).

6.2. *Nigeria 2003*

The Nigerian 2003 presidential elections were selected by Beber and Scacco (2012) as strongly alleged of fraud, and the inspected returns (from the Plateau state) were flagged accordingly by their diagnostics. I reanalyze them classified by numeral, party, and polling station result. Section A.3.2 in the Appendix reports the findings in detail. Under the strong distributional assumption the latent class diagnostics lead to a similar substantive assessment as NHST. Under the relaxed distributional assumption it appears that fraud contaminated all inspected party-result subsets of digits roughly equally, since all are relatively close to a common distribution. Considering other evidence on the election presented by Beber and Scacco (2012) this suggests that all inspected returns were affected by fraud and/or clerical errors to a similar extent.

6.3. *Senegal 2000 and 2007*

The Senegalese presidential elections of 2000 and 2007 differ in that the former are reported to be largely free of fraud, and the latter marred with it (Beber and Scacco, 2012). This makes them an especially interesting test case for inspection under the relaxed distributional assumption–if this assumption holds and the reports are accurate, then digit distributions should vary across elections.

[Table 2 about here.]

Beber and Scacco (2012) pool the numerals into those from winner's (A. Wade) returns and those from selected other columns, and retain the hypothesis of no fraud for 2000 and reject it for 2007. I inspect also the returns of the runners up (A. Diouf in 2000 and I. Seck in 2007). Table 2 reports the digits by numeral, candidate, and year (information on ward result is not provided by Beber and Scacco (2012)).

[Table 3 about here.]

The fit of uniformity to the inspected subsets of returns is reported in Table 3. The χ^2 test rejects at the 5% level the null for winner's last digits in 2000, but not in 2007. However, the sample is much larger in 2007, and the test is sensitive to sample size–if the winner's digits would in 2000 have the 2007 sample size the null would be rejected at the 5% level, and would be retained if in 2007 they would have the 2000 sample size.[6] For the returns of the runners up–not inspected by Beber and Scacco (2012)–uniformity is rejected by the χ^2 test at the 1% level in both elections. A somewhat different picture is provided by the π^* and Δ indexes–the distances from uniformity were similar for the winner in both elections (π^* of 7% vs. 9% and Δ of 2% vs. 3%) and for the runner up of 2000 and the second and third candidate of 2007 (π^* of 16%, 18%, and 16%, and Δ of 4%, 5%, and 5%), respectively. These findings can be interpreted in at least two ways. First, assuming uniformity characterizes fraud-free results, fraud appears similarly prevalent in both elections in the returns for the top two candidates. Alternatively, one can abandon this assumption for the case of Senegal in the observed period.

Under the relaxed distributional assumption, if fraud was relatively low in 2000, but much more present in 2007, then the digit distributions should be different for the two elections, even if the uniform distribution does not characterize the fraud-free last digits. This can be tested with a series of loglinear models reported in Table 4, starting with the independence model. The second model allows the candidates (defined by their placement) to carry different number of territories across elections, imposing the same probability distribution of numerals on all year-candidate combinations. The following three models allow additional interactions.

[Table 4 about here.]

The second model is the simplest with near perfect fit–only only 1% of the observations need to be reallocated or 4% removed for perfect it, as shown in Figure 2. Under the relaxed

[6] For the 2000 density with the 2007 sample size χ^2 is 17 and p-value 0.049 and for the 2007 density with the 2000 sample size χ^2 is 14 and p-value 0.121 (one million simulations from the reference distribution).

distributional assumption both elections seem comparably fraudulent. How to negotiate this inference with the reports of high prevalence of fraud in the second election, but of low one in the first? This might be a fruitful venue for the reanalysis of the reports–it is possible that they underestimate the extent of fraud in 2000 and/or overestimate it in 2007. However, a simpler and more troubling answer is readily available. Namely that–at least for the case of Senegal in the observed period–fraudulent and fraud-free electoral returns are characterized by practically identical probability distributions of last digits. In other words, that the distribution of last digits is not informative with regards to the presence of fraud. This of course means calling into question the enterprise of digit-based election forensics.

[Figure 2 about here.]

7. LIMITATIONS

The value of the empirical demonstration of the proposed solutions to the problems of digit-based election forensics depends on the degree to which the inspected electoral returns represent typical fraudulent and fraud-free elections. This degree has at best a very large uncertainty attached. For that reason, it is common in the literature to validate the methods also using simulations (see Mebane, 2006a; Deckert et al., 2011; Beber and Scacco, 2012). As discussed in Section 2, the value of this is not known. I report the performance of π^* and Δ in such simulations in the Appendix, and do not claim them to be a validation of the methods. Moreover, even if a DBEF method would be convincingly shown to be valid for past elections, it can be invalidated by deliberate behavior on the part of the fraudsters. In case they would deem it worth the costs, the fraudsters can adopt a variety of simple tools that will allow them to fabricate numbers with any digit distribution they desire.

8. CONCLUSION

Among the methods for electoral integrity evaluation digit-based election forensics differentiates itself with its promise to inexpensively evaluate elections requiring only the relevant vote counts. This promise might seem too good to be true, and indeed faces serious challenges. The present paper focuses on whether it can deliver on its promise, and prioritizes solutions that can resolve some of the challenges over the question how well does DBEF perform compared to its alternatives.

Digit-based election forensics relies on the strong assumption that a known probability distribution describes digits in the absence of fraud, but not in its presence. Typically, it evaluates elections with a test of statistical significance of null hypothesis of no fraud based on this assumption. Two independent sets of issues related to the strong distributional assumption and the use of null hypothesis significance testing decrease the usefulness of DBEF. The present paper proposes an alternative to NHST free of its issues, and an independent approach that allows to relax the distributional assumption in some contexts. The former is based on decomposing digit distributions into fraudulent and fraud-free components, and the later on comparing multiple sets of digits using loglinear models.

Reanalysis of Beber and Scacco's (2012) data finds that in two instances the proposed methods lead to similar substantive conclusions, based however on less restrictive assumptions. The finding from the final and arguably the most interesting test are different, and can be interpreted as going against other evidence on the case, or as evidence that even the relaxed distributional assumption is inadequate. The second interpretation suggests that the enterprise of digit-based election forensics is not feasible, since digits can easily be distributed the same way in fraudulent and fraud-free results even if the fraudsters do not deliberately attempt this. Choosing between these two interpretations requires additional information. Given that fair elections are crucial for democratic legitimacy, this can be read as a note of caution for the use of digit-based election forensics.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis, 2nd ed.* John Wiley & Sons.

Alvarez, R. M., L. R. Atkeson, and T. E. Hall (2012). *Evaluating Elections: A Handbook of Methods and Standards*. Cambridge University Press.

Beber, B. and A. Scacco (2011). Replication data for: What the numbers say: A digit-based test for election fraud. `http://hdl.handle.net/1902.1/17151`. Harvard Dataverse, V2.

Beber, B. and A. Scacco (2012). What the numbers say: A digit-based test for election fraud. *Political Analysis 20*(2), 211–234.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572.

Breunig, C. and A. Goerres (2011). Searching for electoral irregularities in an established democracy: Applying Benford's law tests to Bundestag elections in unified Germany. *Electoral Studies 30*(3), 534–545.

Buttorf, G. (2008). Detecting fraud in America's gilded age. Unpublished manuscript, University of Iowa.

Cantú, F. and S. M. Saiegh (2011). Fraudulent democracy? An analysis of Argentina's infamous decade using supervised machine learning. *Political Analysis 19*(4), 409–433.

Clogg, C., T. Rudas, and L. Xi (1995). A new index of structure for the analysis of models for mobility tables and other cross-classifications. *Sociological Methodology*, 197–222.

Clogg, C. C., T. Rudas, and S. Matthews (1997). Analysis of contingency tables using graphical displays based on the mixture index of fit. In J. Blasius and M. Greenacre (Eds.), *Visualization of Categorical Data*, pp. 425–439. Academic Press.

Dayton, C. M. (2003). Applications and computational strategies for the two-point mixture index of fit. *British Journal of Mathematical and Statistical Psychology 56*(1), 1–13.

Deckert, J., M. Myagkov, and P. C. Ordeshook (2011). Benford's law and the detection of election fraud. *Political Analysis 19*(3), 245–268.

Formann, A. K. (2000). Rater agreement and the generalized Rudas–Clogg–Lindsay index of fit. *Statistics in Medicine 19*(14), 1881–1888.

Formann, A. K. (2003a). Latent class model diagnosis from a frequentist point of view. *Biometrics 59*(1), 189–196.

Formann, A. K. (2003b). Latent class model diagnostics–a review and some proposals. *Computational Statistics & Data Analysis 41*(3), 549–559.

Formann, A. K. (2006). Testing the Rasch model by means of the mixture fit index. *British Journal of Mathematical and Statistical Psychology 59*(1), 89–95.

Giles, D. E. (2007). Benford's law and naturally occurring prices in certain ebay auctions. *Applied Economics Letters 14*(3), 157–161.

Gini, C. (1914). Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazione statistiche. *Atti del Reale Instituto Veneto di Scienze, Lettere ed Arti (Series 8) 74*, 185–213.

Hernández, J. M., V. J. Rubio, J. Revuelta, and J. Santacreu (2006). A procedure for estimating intrasubject behavior consistency. *Educational and Psychological Measurement 66*(3), 417–434.

Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical Science*, 354–363.

Ispány, M. and E. Verdes (2014). On the robustness of mixture index of fit. *Journal of Mathematical Sciences 200*(4), 432–440.

Jiménez, R. and M. Hidalgo (2014). Forensic analysis of Venezuelan elections during the Chávez presidency. *PloS ONE 9*(6), e100884.

Judge, G. and L. Schechter (2009). Detecting problems in survey data using Benford's law. *Journal of Human Resources 44*(1), 1–24.

Leemis, L. M., B. W. Schmeiser, and D. L. Evans (2000). Survival distributions satisfying Benford's law. *The American Statistician 54*(4), 236–241.

Mebane, W. R. (2006a). Election forensics: the second-digit Benford's law test and recent American presidential elections. In *Election Fraud Conference*.

Mebane, W. R. (2006b). Election forensics: Vote counts and Benford's law. In *Summer Meeting of the Political Methodology Society, UC-Davis, July*.

Mebane, W. R. (2007). Election forensics: Statistical interventions in election controversies. In *Annual Meeting of the American Political Science Association*.

Mebane, W. R. (2008). Election forensics: Outlier and digit tests in America and Russia. In *American Electoral Process conference, Center for the Study of Democratic Politics, Princeton University*.

Mebane, W. R. (2010a). Election fraud or strategic voting? Can second-digit tests tell the difference? In *Summer Meeting of the Political Methodology Society, University of Iowa*.

Mebane, W. R. (2010b). Fraud in the 2009 presidential election in Iran? *Chance 23*(1), 6–15.

Mebane, W. R. (2011). Comment on "Benford's law and the detection of election fraud". *Political Analysis 19*(3), 269–272.

Mebane, W. R. and K. Kalinin (2009). Comparative election fraud detection. In *Annual Meeting of the American Political Science Association*.

Medzihorsky, J. (2015a). pistar: Rudas, Clogg and Lindsay mixture index of fit. `https://github.com/jmedzihorsky/pistar`. R package version 0.5.2.5.

Medzihorsky, J. (2015b). Replication data for: Election fraud: A latent class framework for digit-based tests. http://dx.doi.org/10.7910/DVN/1FYXUJ. Harvard Dataverse, V1 [UNF:6:FIWHvsHNzZgPStT0+kgbsQ==].

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics 4*(1), 39–40.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review 109*(2), 330.

Norris, P., R. W. Frank, and F. M. i Coma (2014). *Advancing Electoral Integrity*. Oxford University Press.

Pericchi, L. and D. Torres (2011). Quick anomaly detection by the Newcomb–Benford law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. *Statistical Science 26*(4), 502–516.

Revuelta, J. (2008). Estimating the π^* goodness of fit index for finite mixtures of item response models. *British Journal of Mathematical and Statistical Psychology 61*(1), 93–113.

Rudas, T. (1998). The mixture index of fit. In A. Ferligoj (Ed.), *Advances in Methodology, Data Analysis, and Statistics*, pp. 15–22. FDV.

Rudas, T. (1999). The mixture index of fit and minimax regression. *Metrika 50*(2), 163–172.

Rudas, T. (2002). A latent class approach to measuring the fit of a statistical model. In J. A. Hagenaars and A. L. McCutcheon (Eds.), *Applied latent class analysis*, pp. 345–365. Cambridge University Press.

Rudas, T. (2005). Mixture models of missing data. *Quality & Quantity 39*(1), 19–36.

Rudas, T., C. Clogg, and B. Lindsay (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(4), 623–639.

Rudas, T. and E. Verdes (2015). Model based analysis of incomplete data using the mixture index of fit. In G. R. Hancock and G. B. Macready (Eds.), *Advances in latent class analysis: A Festschrift in Honor of C. Mitchell Dayton*. Information Age Publishing.

Rudas, T. and R. Zwick (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics 22*(1), 31–45.

Tam Cho, W. K. and B. J. Gaines (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American Statistician 61*(3), 218–223.

Verdes, E. and T. Rudas (2003). The π^* index as a new alternative for assessing goodness of fit of logistic regression. In Y. Haitovsky and Y. Ritov (Eds.), *Foundations of Statistical Inference*, pp. 167–177. Springer.

Ziliak, S. T. and D. N. McCloskey (2008). *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*. University of Michigan Press.

List of Figures

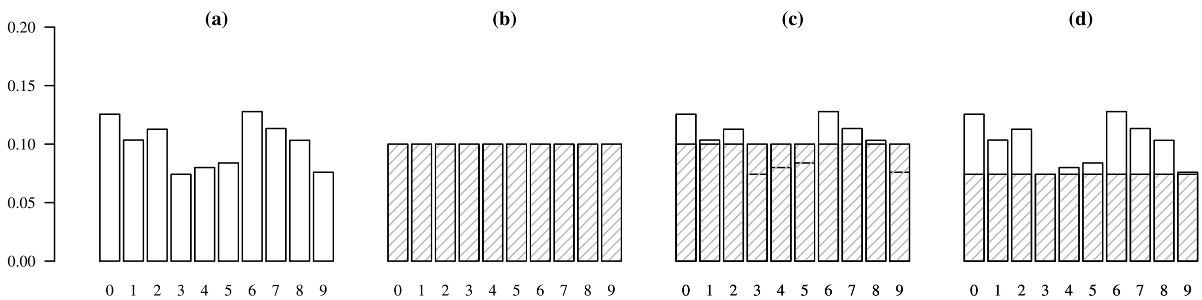

Figure 1: Illustration of the mixture index of fit decomposition of an observed set of digits (a) using a discrete uniform model (b). The highest ratio of the model density over the observed density is for digit '3' (c). The resulting latent class model (d) has a π^* value of 0.26. Simulated data.

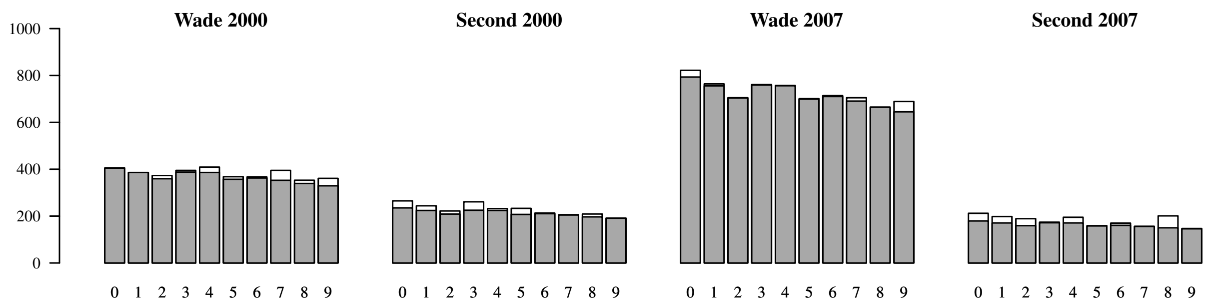

Figure 2: Model fit under π^* of the second loglinear model (year-candidate, numeral) to the Senegalese data. N=15,172. Model distribution in grey, residuals in white.

List of Tables

Table 1: Classification of last digits from ward returns by the numeral and whether the ward has been inspected by election observers.

	Observers	
Numeral	Yes	No
0	$d_{1,1}$	$d_{1,2}$
1	$d_{2,1}$	$d_{2,2}$
2	$d_{3,1}$	$d_{3,2}$
3	$d_{4,1}$	$d_{4,2}$
4	$d_{5,1}$	$d_{5,2}$
5	$d_{6,1}$	$d_{6,2}$
6	$d_{7,1}$	$d_{7,2}$
7	$d_{8,1}$	$d_{8,2}$
8	$d_{9,1}$	$d_{9,2}$
9	$d_{10,1}$	$d_{10,2}$

Table 2: Last digits in election results in Senegalese presidential elections of 2000 and 2007 for the winner in both (Wade) and the candidate that placed second (Diouf in 2000 and Seck in 2007). Only numbers with three or more digits included. N=15,172. Source: author's calculation. Data source: Beber and Scacco (2012).

	2000		2007	
	1st	2nd	1st	2nd
0	405	265	822	212
1	386	244	764	198
2	373	222	705	189
3	395	261	761	174
4	409	232	757	195
5	368	233	701	159
6	367	213	714	170
7	395	206	705	156
8	353	209	665	201
9	361	191	689	147

Table 3: Fit of the uniform distribution to ten subsets of last digits in the Senegalese electoral returns. Reference distributions of the test statistics obtained with one million simulations. Fraction sizes in %. Jackknifed confidence intervals for π^* and Δ. 'Other col.' refers to returns pooled by Beber and Scacco (2012).

		N	χ^2	p	π^*	95% ci	Δ	95% ci
2000	1st (Wade)	3,812	8.89	0.45	7	(0, 17)	2	(1, 4)
	2nd (Diouf)	2,276	22.97	0.01	16	(5, 27)	4	(2, 6)
	First Two	6,088	23.02	0.01	9	(2, 17)	3	(1, 4)
	Registered	8,058	7.18	0.62	5	(0, 11)	1	(0, 2)
	Other Col.	10,334	5.34	0.80	3	(0, 9)	1	(0, 2)
2007	1st (Wade)	7,283	26.82	<0.01	9	(2, 15)	3	(1, 4)
	2nd (Seck)	1,801	24.08	<0.01	18	(6, 31)	5	(3, 8)
	First Two	9,084	37.09	<0.01	8	(2, 14)	3	(2, 4)
	3rd (Dieng)	1,091	15.15	0.09	16	(0, 32)	5	(2, 8)
	Other Col.	2,892	26.54	<0.01	15	(5, 25)	4	(2, 6)

Table 4: Model fit for five loglinear models fit to the Senegalese data. N=15,172. Fraction sizes in %. Jackknifed confidence intervals.

	χ^2	df	p	π^*	95% ci		Δ	95% ci	
Independence	596.73	28	<0.01	12	(10,	14)	8	(8,	9)
Candidate-Year, Numeral	26.53	27	0.49	4	(0,	8)	1	(1,	2)
Candidate-Year, Numeral-Candidate	13.25	18	0.78	2	(0,	4)	1	(0,	2)
Candidate-Year, Numeral-Year	23.13	18	0.19	3	(1,	4)	1	(1,	2)
Candidate-Year, Numeral-Year, Numeral-Candidate	8.39	9	0.50	1	(0,	3)	1	(0,	1)