# Supplemental Materials for: An Informed Forensics Approach to Detecting Vote Irregularities

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

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

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

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

# 0 APPENDIX A: ELECTION COVERAGE

**Table A1.** Countries and elections included in analysis

| **Country** | **Years covered** | **Country** | **Years covered** |
|---|---|---|---|
| Albania | 2001-2005 | Italy | 1948-2006 |
| Argentina | 1983-2003 | Jamaica | 1967-2002 |
| Australia | 1946-2001 | Japan | 1947-1993 |
| Austria | 1945-2008 | Kenya | 1992-1997 |
| Bangladesh | 1973, 2001-2008 | Latvia | 1993-2006 |
| Belgium | 1946-2007 | Lithuania | 1992-2008 |
| Bolivia | 1985-2005 | Malawi | 1999-2004 |
| Bosnia & Herzegovina | 1996-2006 | Mauritius | 1995-2000 |
| Botswana | 1969-2004 | Mexico | 1991-2006 |
| Brazil | 1947-1962, 1982-2006 | Moldova | 1994-1998 |
| Bulgaria | 1991-2005 | Netherlands | 1952-2006 |
| Cambodia | 2008 | New Zealand | 1946-1999 |
| Cameroon | 1997-2002 | Niger | 1999 |
| Canada | 1945-2008 | Nigeria | 2003 |
| Colombia | 1958-2006 | Norway | 1945-2005 |
| Costa Rica | 1953-2006 | Pakistan | 2002-2008 |
| Croatia | 2000-2007 | Philippines | 1992-2010 |
| Cyprus | 1981-1996 | Poland | 1991-2007 |
| Czechoslovakia | 1990-1992 | Portugal | 1976-2005 |
| Czech Republic | 1998-2006 | Romania | 1990-2004 |
| Denmark | 1945-1998 | Singapore | 1968-2006 |
| Dominican Republic | 1962-2006 | Slovakia | 1994 |
| Equatorial Guinea | 1993 | Slovenia | 1996-2008 |
| Estonia | 1992-2007 | Spain | 1977-2008 |
| Finland | 1945-2007 | South Africa | 1994-1999 |
| France | 1973-2007 | Sri Lanka | 1952-1977, 1989-2010 |
| Gambia | 1997, 2007 | Sweden | 1948-2006 |
| Ghana | 1996-2004 | Switzerland | 1947-2007 |
| Greece | 1946-2007 | Taiwan | 1986-2004 |
| Guinea Bissau | 1994, 2004 | Thailand | 1969-1992 |
| Guyana | 1997-2006 | Trinidad & Tobago | 1966-2002 |
| Honduras | 1981-2001 | Turkey | 1961-2002 |
| Hungary | 1990-2001 | United Kingdom | 1945-2005 |
| Indonesia | 1999-2004 | United States | 1946-2006 |
| Ireland | 1948-1997 | Venezuela | 1958-1988 |
| Israel | 1949-2009 | Zambia | 1968, 1991-2006 |

# 0 APPENDIX B: CODING

Our measure of the extent of inequality is the Gini Index at the national level, which we take from the Luxembourg Income Study, the World Income Inequality Database, and the Measuring Income Inequality Database. Our measure of ethnolinguistic fractionalization is taken from Fearon (2003). Data on the percentage of the population living in *urban* centers was collected from the United Nations' World Urbanization Prospects (2011 Revision). The reports provide percentages of the population in a country living urbanely at five-year increments Between-years were imputed by calculating yearly linear incremental change between each five-year measurement. Where possible, figures for *average district magnitude* were calculated manually from data reported by the Global Elections Database and CLEA dataset and, where these figures were not reported, we augmented them with data reported by Golder (2005). We collected data on *turnout* from the International Institute for Democracy and Electoral Assistance *Voter Turnout Database*.

In seven cases, the system duration variable was missing for democratic regimes. These cases were coded as new democracies in the results below. Finally, both the system duration and polity measures were missing for 13 cases, and these observations were placed in a distinct missing category.[1]

The NELDA indicators we used to construct our measure of fraud are shown in Figure B1. We combined these indicators using a standard item response theoretic (IRT) model (Baker and Kim, 2004) to generate a latent – but explicitly indicative – *IRT fraud score* for each observation. Because the NELDA dataset does not distinguish among districts or regions

**Figure B1.** Items used to construct our measure of fraud

1. Before elections, are there significant concerns that elections will not be free and fair? (nelda11)
2. Were there riots or protests *and* did those riots or protests involve allegations of fraud? (nelda30)
3. Were results that did not favor the incumbent canceled? (nelda32)
4. Were results that were favorable to the incumbent canceled? (nelda34)
5. If western monitors were present, were there allegations by Western monitors of significant vote-fraud? (nelda47)
6. Were some election monitors denied the opportunity to be present by the government holding elections? (nelda48)
7. Did any monitors refuse to go to an election because they believed that it would not be free and fair? (nelda49)

[1]The BART approach easily handles missing data, so long as it is nominal. For nominal variables, we simply add a category for "missing", which the BART model then treats as just one more (unordered) category. While some information is lost in this categorization, this approach is still far superior to simple listwise deletion. To our knowledge, alternative strategies such as multiple imputation have not been implemented within a BART framework. We followed this categorization procedure for four variables: average district magnitude, economic inequality, GDP growth, and political regime.

**Table B1.** Coding details for forensic indicators and contextual risk factors

| **Variable** | **Operationalization** |
|---|---|
| *Distance of Last Two Digits* | *Ave. difference between last and second-to-last digits in vote totals* |
| Continuous Over [0, 9] | Min: 2.38 Max: 4.15 Mean: 3.11 |
| *Final Digit (Uniform Violation)* | $\chi^2$ *goodness-of-fit test statistic of relative frequencies of numbers in the last digit position (expected to be uniformly distributed)* |
| Continuous over [0, $\infty$] | Min: 5.948E-4 Max: 1.43 Mean: 0.09 |
| *Second Digit (Benford Violation)* | $\chi^2$ *goodness-of-fit test statistic of relative frequencies of numbers in the second digit position (expected to be Benford distributed)* |
| Continuous over [0, $\infty$] | Min: 8.968E-4 Max: 1.72 Mean: 0.10 |
| *Second Digit (Mean)* | *Mean of numbers in the second digit position* |
| Continuous over [0, 9] | Min: 3.0 Max: 6.3 Mean: 4.19 |
| *Economic Inequality* | *Gini coefficient, discretized into quartiles* |
| First Quartile | Gini Values from: 20.13 To: 27.30 |
| Second Quartile | Gini Values from: 27.31 To: 32.70 |
| Third Quartile | Gini Values from: 32.71 To: 43.00 |
| Fourth Quartile | Gini Values from: 43.01 To: 60.10 |
| *Ethnic Fractionalization* | *Fearon's index of fractionalization* |
| Continuous Over [0, 1] | Min: 0.01 Max: 0.89 Mean: 0.33 |
| *Urban Population* | *Percent of population living in urban centers* |
| Continuous Over [0, 100] | Min: 7.04 Max: 100.00 Mean: 63.79 |
| *Average District Magnitude* | *Mean of all districts' magnitudes, discretized into quartiles* |
| First Quartile | Mean Magnitudes from: 1.00 To: 1.00 |
| Second Quartile | Mean Magnitudes from: 1.01 To: 6.17 |
| Third Quartile | Mean Magnitudes from: 6.16 To: 11.10 |
| Fourth Quartile | Mean Magnitudes from: 11.11 To: 150.00 |
| *National Turnout* | *Percent registered voters casting ballots in election* |
| Continuous Over [0, 100] | Min: 2.73 Max: 99.41 Mean: 75.38 |
| *Regime Type* | *Average score from prior election, discretized into regime type* |
| Autocracy | Polity values from: -10 To: -6 |
| Anocracy | Polity values from: -5 To: 5 |
| New Democracy | Polity value over 4 for less than 10 years |
| Old Democracy | Polity value over 4 for over 10 years |
| *GDP Change* | *Change in GDP per capita in the prior year* |
| First Quartile | Values from: -32.1 To: 1.81 |
| Second Quartile | Values from: 1.81 To: 3.56 |
| Third Quartile | Values from: 3.56 To: 4.46 |
| Fourth Quartile | Values from: 5.58 To: 34.8 |
| *Regime Crisis* | *Coups, revolutions, state failures, and fractioning* |
| | Stable: 82.8% Unstable: 17.2% |
| *Independent Commission* | *Level of independence of electoral commission* |
| Government | Government-run: 38.6% Mixed: 47.8% Independent: 13.5% |

*within* a country, our measure of fraud is available only at the country-election level. That is, the score is assigned to an election as a whole rather than specific results in geographic subunits.

The parameter estimate for the IRT model are shown in Table B2.[2] These estimates were generated using the full NELDA dataset (n=1186). All items loaded significantly on the underlying factor with the largest discrimination parameters (Items 5 and 6) related to reported irregularities by election monitors. The smallest discrimination parameters (Items 3 and 4) related to canceled elections, which are conceptually less directly related to election irregularities.[3]

**Table B2.** IRT model for our measure of fraud

| | Difficulty | Discrimination |
|---|---|---|
| Item 1: Concerns | 0.410 | 1.684 |
| | (0.049) | (0.209) |
| Item 2: Riots | 3.759 | 1.440 |
| | (0.573) | (0.316) |
| Item 3: Canceled, unfavorable | 3.746 | 1.391 |
| | (0.571) | (0.301) |
| Item 4: Canceled, favorable | 2.229 | 1.386 |
| | (0.191) | (0.181) |
| Item 5: Monitor allegations | 2.152 | 2.251 |
| | (0.157) | (0.369) |
| Item 6: Monitors denied | 2.022 | 3.394 |
| | (0.125) | (0.782) |
| Item 7: Monitors refused | 2.158 | 1.865 |
| | (0.156) | (0.252) |
| n | 1,816 | |
| BIC | 5650 | |
| $\alpha$ | 0.509 | |

Standard errors in parentheses.

Much like other covert activities, such as political corruption, directly assessing fraud is highly difficult, if not altogether impossible. Because of this, of course, we must admit that some of our IRT items are not directly indicative of objective, empirically-measurable fraud. Our strategy in constructing the IRT measure of fraud came down to assessing expectations of fraud, implications of fraud, and allegations of fraud that all also exhibited a high level of internal consistency. That is, our battery is substantively informed, but it is also statistically verifiable that these particular indicators point in the same direction along our latent variable of interest.

[2] This model was fit using the `ltm()` command provided by the `ltm` package for R version 3.0.1.

[3] As we would expect, none of these indicators occur regularly. The proportion of observations in the full dataset ($n = 1,816$) that meet each criteria are 0.3871 (Concerns), 0.0116 (Riots), 0.0132 (Canceled unfavorable), 0.083 (Cancelled favorable), 0.0463 (Monitor allegations), 0.039 (Monitors denied), 0.0595 (Monitors refused).

By way of substantively justifying our selection of items from the NELDA database, we turned to other prominent sources of data on fraud, in particular the Quality of Elections Database (QED) managed by Judith Kelley and the Election Integrity Project managed by Pippa Norris and her several collaborators. Both of these data repositories focus on indicators that are similar to the ones we employ, especially questions about pre-election monitoring and pre-election expectations about outcomes and fraudulent activity. We drew, in particular, on Kelley's data which has greater cross-sectional and temporal scope.

Kelley's database works with coded observer reports issued from several election monitoring agencies. The QED addresses numerous subjective and objective components of fraudulent behaviors, among them some of the same indicators (or similar indicators) that we employ: measures of pre-election political conditions and pre-election administrative irregularities; tracking the number of pre-election assessment visits from monitoring organizations; tracking the number of press statements issued before the election by these organizations; and assessing the extent of the invalidation or cancellation of results after the election. Ultimately, then, it seems that something of a consensus exists across major databases that set out to measure fraud: namely, that the pre-election expectations game and the post-election invalidation of results are both signals of election manipulation motivated by less-than-democratic intent on behalf of political elites.

A final issue with our selection of indicators is the extent to which "free and fair" maps into "not fraudulent" while "unfree" or "unfair" maps into "fraud." The distinction here is mainly rhetorical and boils down to the ways in which different strains of the comparative literature think about this particular phrasing. In the context of cross-national regime studies that rely on data like, for example, Polity IV, we acknowledge that the designation of "free and fair" would not translate very readily into "not fraudulent" because there are many more dimensions of the former concept than the latter. However, specifically related to studies of fraud and election monitoring, the rhetoric of "free and fair" tends to be invoked as an indicator of whether or not an election's results straightforwardly "represented the will of the people" or not (see extensive discussion in the QED codebooks provided by Judith Kelley). Because NELDA, the QED, and even Sarah Birch's Electoral Malpractice databases all rely to some extent on these election observation reports, their use of "free and fair" can be understood to be indicative of an absence of pre- or post-election tampering with the results of the balloting process.

# 0 APPENDIX C: VALIDATION OF OUR FRAUD MEASURE

The database on electoral malpractice first developed for Central and Eastern European countries in Birch (2007) was subsequently extended to a handful of additional countries in Latin America and Africa in Birch (2012). Birch draws on election observation reports from the Organization for Security and Co-operation in Europe (OSCE), which conducts monitoring missions with the intent of diagnosing the extent to which a given country is able to administer a "clean" election in line with OCSE criteria. For each mission, the organization then subsequently published an online assessment of what the observers witnessed firsthand on the ground. Birch relies on three separate coders to assign scores ranging from 1 (an election substantially in compliance with OSCE criteria) to 5 (an election where the criteria were substantially violated). The scale, then, is a measure of increasingly systematic and problematic malpractice.

The *Quality of Elections* database by Kelley (2012) also hand-codes observer reports on election quality, but from a different source: "Country Reports on Human Rights Practices" published as an annual report by the U.S. State Department. Relevant for our purposes, Kelley codes for both the *severity* and the *prevalence* of election irregularities, which include many of the aspects of fraud that our IRT model is intended to register: vote padding, inflated vote counts, ballot stuffing, problems in the counting or tabulation of votes, etc. The resulting ordinal metrics both range from 0 (no problems with cheating) up to 3 (major and systemic problems along the lines described above). We combine these two indicators linearly to capture a more nuanced, 6-point scale which allows us to capture variation in assessments along the conceptual lines of "no problems" to "small problems, sporadic" to "severe problems, sporadic" to "severe problems, systemic."

Despite their many merits, relative to our data set, both the electoral malpractice and *QED* databases come with fairly significant geographical and selection biases in that (a) their coverage of cases and elections is largely confined to Central and Eastern Europe, Latin America, Africa and – in the case of the *QED* – a handful of cases from the Middle East and Asia; and (b) by virtue of drawing on election monitoring reports, they may be focusing on countries where the *a priori* suspicion of fraud was high.

Despite these facts, the two databases are among the best objective (more-or-less) context-rich indicators of fraud. We think it is a productive exercise, then, for those countries and elections that overlap across data sets, to compare our IRT scores against these two metrics. If we observe a high level of coherence between our tool's assessment and these assessments for the limited set of cases held in common across the data sets (or, in the absence of this, if we can reasonably account for discrepancies), then we can extend our IRT scoring into other countries and years with a high degree of confidence in the validity of our tool.[4] What we end up finding is that our IRT score correlates quite well with either of these other metrics. This discussion is included in the main text of the manuscript.

[4]By virtue of both drawing on election monitoring reports as their informational inputs, it is not terribly surprising that Birch's and Kelley's scores correlate relatively highly with one another.

# 0 APPENDIX D: FIT STATISTICS

Three indicators, root mean squared error (RMSE) and median absolute deviation (MAD), and mean absolute error (MAE) provide summaries for the error of each model. Median absolute percentage error (MEAPE), on the other hand, measures error as a proportion of the dependent variable.

Denote the prediction for some observation $i$ as $p_i$ and the observed outcome as $y_i$. We define the *absolute error* as $e_i \equiv |p_i - y_i|$ and the *absolute percentage error* as $a_i \equiv e_i/|y_i| \times 100$. Denoting the median of some vector $\mathbf{x}$ as $med(\mathbf{x})$, we define the following statistics:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{\sum_1^n e_i^2}{n}} \\ \text{MAD} &= \text{med}(\mathbf{e}) \\ \text{MAE} &= \frac{\sum_1^n e_i}{n} \\ \text{MAPE} &= \frac{\sum_1^n a_i}{n} \\ \text{MEAPE} &= \text{med}(\mathbf{a}) \end{aligned}$$

# 0 APPENDIX E: GENERALIZATION ERROR

To alleviate concerns about the generalizability of our results, we rely on two approaches to estimate out-of-sample (or generalization) error. First, we randomly partition the entire data into 15 subsets and evaluate each model by the fit statistics above. In this 15-fold cross-validation, we reserve approximately $14/15^{ths}$ of the data for training BART model, and test the accuracy of the model against the remaining $1/15^{th}$. The results suggest that the informed forensics approach – broadly speaking – outperforms the other models, although it is clear that the contextual risk factors by themselves are able to predict fraudulent elections out-of-sample quite well. Specifically, the consensus rank for the informed forensics model was lowest (best) in 9 out of the 15 partitions. However, *in all cases* the consensus rank for the informed model is lower than for the pure forensics model. More generally, the mean consensus rank[5] for the informed forensics, forensics only and contextual only models were 1.4, 2.85 and 1.75 respectively across the 15 partitions. In all, these results support our claim that the informed forensics approach to detecting fraud provides an efficient way for supplementing forensic tools in identifying instances of electoral fraud as evaluated by out-of-sample predictions.

Second, we calculate the *leave-one-out bootstrap error* designed by Efron and Tibshirani (1997), using 150 bootstrap samples. In this case, and to account for the fact that we observe data from contiguous elections in the same country, we use the moving block method (Künsch, 1989) to create the bootstrap samples, where we define overlapping blocks that correspond to a set of three sequential elections from the same country. By sampling at the block level, our aim is to generate estimates that are more robust to violations of the assumption of independent errors resulting from including elections close in time and from the same country.

For each observation $i \in [1, \ldots, n]$, we first calculate the average loss across all bootstrap samples where observation $i$ is *excluded* from the training sample in order to avoid testing a model with observations used to train it. Efron and Tibshirani (1997) suggest then taking the mean of this quantity for all observations,[6] While we originally also include measures that take the median over all observations, an expression for the standard error of this estimate is available only for the mean estimate (Efron and Tibshirani, 1997). The results are shown in Table E1.[7] The results are generally consistent with our findings from the 15-fold cross-validation study. That is, in the aggregate the informed forensics approach generally does better although there is clearly a great deal of variation across bootstrap samples and observations as to which model does best. While the informed forensics model does best, the wide standard errors indicate that it is not obviously and everywhere dominant. However, we again caution that these results should not be weighted too strongly as bootstrap and cross-validation approaches to calculating test errors rates are not recommended for tree-based models.

---

[5] This is calculated as the mean of the consensus ranks across the 15 random folds.

[6] Specifically, we calculate the mean loss across all bootstraps that exclude observation $i$

[7] Once again, and for each bootstrap sample, we allowed BART to run for 50,000 total iterations and used only the final 5,000 iterations for our analyses.

**Table E1.** Leave-one-out bootstrap generalization error

| | RMSE | MAE | MAPE | Consensus Rank |
|---|---|---|---|---|
| Informed Forensics | 0.2077 | 0.2376 | 0.4682 | 1.67 |
| | (0.07) | (0.05) | | (0.11) |
| Forensic Tools Only | 0.1972 | 0.2656 | 0.5057 | 2.33 |
| | (0.08) | (0.04) | | (0.08) |
| Contextual Risks Only | 0.2080 | 0.2369 | 0.4689 | 2.00 |
| | (0.07) | (0.04) | | (0.10) |

Generalization error estimates generated using 150 bootstraps (Efron and Tibshirani, 1997). Bootstraps samples are created using overlapping within-country blocking to control for non-independence. Quasi-standard errors are in parentheses.

# 0 APPENDIX F: ADDITIONAL DISCUSSION OF BART

Conceptually, it is easiest to think of the BART model as a method for creating an ensemble of so-called "weak learners." In recent years, such ensemble methods have come to play a leading role in the machine-learning and nonparametric statistics community (Hastie, Tibshirani and Friedman, 2009). A wide range of approaches, including neural nets (King and Zeng, 2001), ensemble Bayesian model averaging (Montgomery, Hollenbach and Ward, 2012), $k$-nearest neighbors, and more can be conceptualized as variations on the ensemble approach. Of particular relevance here is the success of boosting (Freund and Schapire, 1997; Friedman, 2001), bagging (Breiman, 1996), and random forests (Breiman, 2001), which are all variants of tree-based algorithms. These approaches to classification and prediction have been advertised as the "best off-the-shelf classifier[s] in the world" (Zhu et al., 2009, 350), and are equally powerful in prediction tasks. However, in addition to out-performing these methods in many prediction tasks (Chipman, George and McCulloch, 2010), one crucial advantage of using BART is that Bayesian estimation techniques allow the model to produce measures of uncertainty regarding not only the model's parameters (which are often not of direct interest), but also of more usual quantities of interest, such as the partial dependence of the outcome of interest across any one combination of covariates.

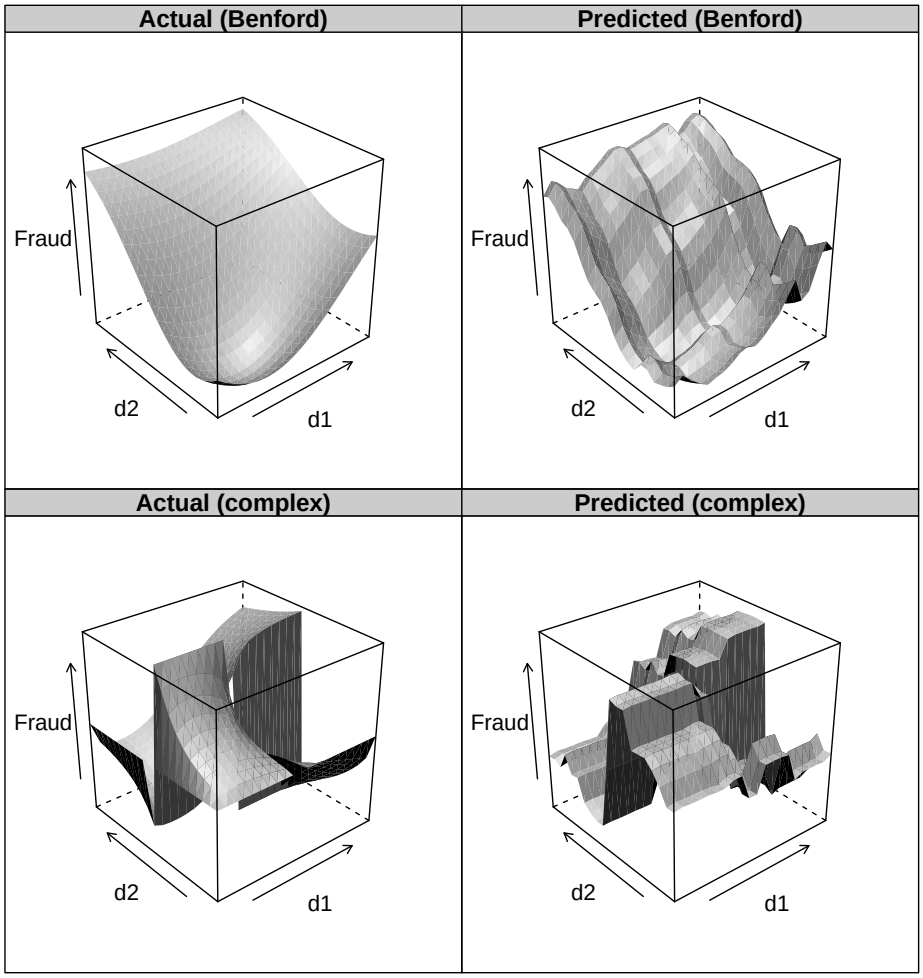
To illustrate some of the favorable properties of the BART method, Figure F1 shows the results of two simulation exercises where some outcome variable (Fraud Proxy) is predicted in a non-linear, interactive manner by two predictive variables (d1 and d2). Conceptually these may be thought of as the frequency of digits in election returns, qualitative risk factors for fraud, or both. The left panels show "true" relationships between the predictors and the outcome while the right panel shows the relationships recovered by the BART model estimated on $n = 350$ randomly generated observations.[8]

[8] The data generating process (DGP) for the top panels is

$$z = 1 - F\left(\frac{\sum_{i\in 1,2}(E_i - x_i)^2}{E_i}\right)$$

where $F(.)$ is the cumulative density function (CDF) of a $\chi^2$ distribution with one degree of freedom, and

**Figure F1.** Simulated and recovered interactive relationships




The left panel shows the true *simulated* relationship between two predictor variables (d1, d2) and an outcome variable (Fraud Proxy). The right panel shows the recovered relationship between these three variables as estimated by a BART model on 350 random observations. The figure shows the BART model is able to correctly recover both smooth (top) and complex (bottom) interactive relationships between predictor variables and some outcome of interest.

The top panels of Figure F1 illustrate a "smooth" interactive relationship between the two predictor variables and the outcome, similar in character to that predicted by violations to the first digit Benford's law. The bottom panels show a much more complex interactive relationship between the predictors and the outcome. In both cases, however, BART is able to recover the relationship based on a relatively small number ($n = 350$) of observations.

# 0 REFERENCES

Baker, Frank B. and Seock-Ho Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

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

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

Breiman, L. 1996. "Bagging predictors." *Machine Learning* 26:123–140.

Breiman, L. 2001. "Random forests." *Machine Learning* 45:5–32.

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

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

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth* 8:195–222.

Freund, Y. and R.E. Schapire. 1997. "A decision-theoretic generalization of online learning and an application to boosting." *J. Comput. System Sci.* 55:119–139.

Friedman, J.H. 2001. "Greedy function approximation: A gradient boosing machine." *Ann. Statist* 29:1189–1232.

Golder, Matt. 2005. "Democratic Electoral Systems Around the World, 1946-2000." *Electoral Studies* 24:103–121.

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

---

$E = \{.301, .176\}$. The DGP for the bottom panels is

$$z = C\left(\sqrt{x_1 \times x_2}\frac{(0.5 - x_1)^2}{0.5(0.5 - x_1)(0.5 - x_2)}\right),$$

where $C(.)$ is the CDF of the Cauchy distribution. The BART models were estimated using the parameter setting of: `ntree=200`, `sigquant=.95`, `sigdf=2`, `k=2`, and `base=.95`. These parameters, which construct the regularization priors, are discussed in greater detail in Chipman, George and McCulloch (2010).

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

King, Gary and Langche Zeng. 2001. "Improving Forecasts of State Failure." *World Politics* 53(4):623–658.

Künsch, Hans R. 1989. "The Jackknife and the Bootstrap for General Stationary Observations." *The Annals of Statistics* 17(3):1217–1241.

Montgomery, Jacob M., Florian Hollenbach and Michael D. Ward. 2012. " Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20(3):271–291.

Zhu, Ji, Hui Zou, Sharon Rosset and Trevor Hastie. 2009. "Multi-class AdaBoost." *Statistics and Its Interface* 2:349–360.