Avoiding Guilt by Association: The Electoral Consequences of Exposure to Nearby Corruption in Brazil

Appendix

Gustavo Diaz*

2020-08-11

A. Coding Audits Before 2006

Protocol

I use text data from the audit reports as a bridge between labeled and unlabeled cases. I use a bag-of-words approach to predict the sum of moderate and severe infractions, divided by the number of service orders. The predictors are raw word counts. I use the following protocol:

1. Match text data from the audit reports with CGU infraction labels for the 2006-2015 period. This is the period where the CGU coding is available.
2. Predictors are word counts, omitting infrequent terms (words missing in more than 99% of the documents).
3. This leaves a data set with 1226 observations and 11386 variables.
4. Randomly split data in training (75%) and test (25%) sets.
5. Fit multiple random forest on training data with a grid of tuning parameters, choose the model and tuning parameters with the lowest RMSE, create predicted variable in test set.

I chose random forests because they achieve reasonable performance with the current data. I explored including topic membership covariates from structural topic modeling to assist the algorithm, but the predictive gains

*PhD Candidate. Department of Political Science. University of Illinois at Urbana-Champaign. E-mail: diazdia2@illinois.edu

are minimal. An alternative is to use algorithms from the deep-learning family, but trial runs suggest that the sample size is too small to guarantee convergence.

One way to increase predictive power dramatically would be to turn this from a regression problem into a classification task by separating documents into findings. That is, moving from predicting numbers of infractions at the document level to predicting whether each item counts as a formal, moderate, or severe infraction. This yields a larger training set with more information, and also supervised learning algorithms tend to perform better with classification tasks than with continuous outcomes. However, because audit report formats are not stable over time, dividing documents at the finding level would require prohibitively expensive human coding.

Performance

Figure A1 reports reports performance in the test set (N = 319). In average, the predicted values are off by 1.34 infractions per service order compared to the actual values. The predictions map close to a 1:1 relationship for moderate cases of corruption, but tend to underestimate it for large outliers. This implies that models using this variable will underestimate the effect of nearby corruption on the outcomes of interest, making it harder to detect non-zero estimates.

Validation

As a validation exercise, I reproduce the findings in previous work using the machine coded categories. Rundlett (2018) shows that exposing corruption has a negative effect on incumbent vote only for the 2004 elections. Table A1 replicates the same pattern using my own data set. This is different from the main analysis in that it evaluates direct effects: Whether revealing corruption in a municipality affects votes for the incumbent in that municipality. The substantive result is the same as in previous work.

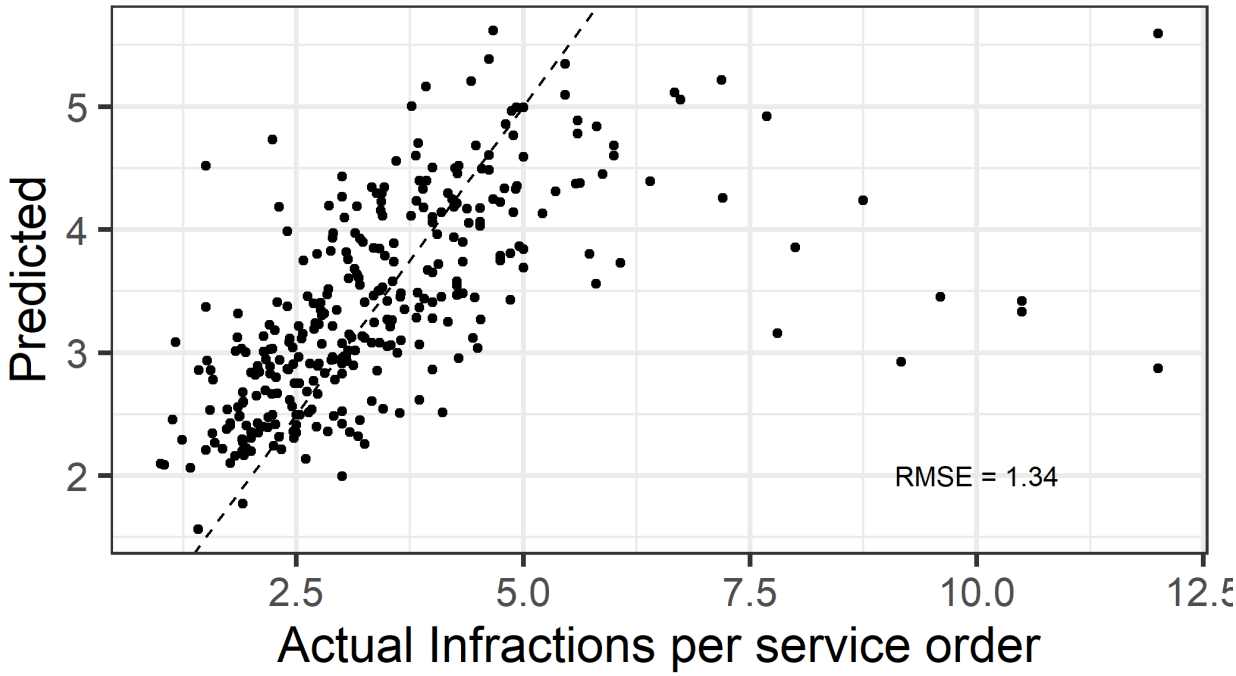

Figure A1: Actual vs. predicted corruption variable in the test set. The dashed line denotes the 1:1 relationship. The algorithm does performs well at predicting moderate levels of corruption, but tends to underestimate large outliers.

	1. Pooled	2. 2004	3. 2008	4. 2012	5. 2016
Intercept		0.63*	0.49*	0.48*	0.44*
		(0.05)	(0.03)	(0.02)	(0.06)
Infractions	−0.00	−0.06*	0.00	0.00	0.00
	(0.00)	(0.02)	(0.01)	(0.01)	(0.02)
R^2	0.01	0.04	0.00	0.00	0.00
Adj. R^2	0.01	0.03	-0.00	-0.00	-0.01
Num. obs.	1152	211	464	341	136
RMSE	0.16	0.14	0.17	0.14	0.21
N Clusters	4				

*$p < 0.05$

Table A1: Replication of Rundlett (2018) with my data. The first column includes election year fixed effects and clustered standard errors by election year. The remaining columns include robust (HC1) standard errors.

B. Regression tables from the results section

The main text reports results using figures. This section shows tables with numerical results that underlie those figures. The list below shows the correspondence:

- Table B1 shows the results of Figure 2 in the main text: The effect of nearby corruption on party switching by audit status using different definitions of nearby.
- Table B2 shows the results in Figure 3: The effect of nearby corruption in interaction with the proportion of same-party audited mayors on party switching. Note that Figure 3 shows simulated marginal effects at discrete margins based on this estimation, the interaction term is indistinguishable from zero.
- Table B3 shows the results in Figure 4: The effect of nearby corruption at the ninth cumulative contiguity order on party switching divided by election year.
- Tables B4 and B5 show the results of panels A and B of Figure 5, respectively: The effect of nearby corruption on seeking and securing reelection.

Contiguity	Infractions	SE	p-value	Audited	SE	p-value	Interaction	SE	p-value	N	Adj. R-squared
1	0.009	0.004	0.117	0.023	0.046	0.651	-0.005	0.016	0.755	6324	0.025
2	0.007	0.002	0.027	0.026	0.010	0.076	-0.005	0.002	0.113	12175	0.021
3	0.004	0.001	0.031	0.003	0.014	0.842	-0.001	0.002	0.654	14566	0.019
4	0.005	0.001	0.025	0.009	0.013	0.532	-0.001	0.002	0.423	15043	0.019
5	0.005	0.001	0.012	0.001	0.012	0.919	-0.000	0.001	0.730	14830	0.020
6	0.005	0.001	0.011	0.017	0.009	0.171	-0.001	0.001	0.196	14441	0.021
7	0.005	0.001	0.010	0.021	0.009	0.092	-0.002	0.001	0.090	13966	0.021
8	0.005	0.001	0.009	0.036	0.005	0.004	-0.002	0.000	0.008	13411	0.020
9	0.004	0.001	0.009	0.029	0.010	0.057	-0.001	0.001	0.103	12755	0.021
10	0.004	0.001	0.010	0.037	0.018	0.128	-0.002	0.001	0.173	12052	0.021

Table B1: The effect of nearby corruption on party switching by audit status using different definitions of nearby.

Contiguity	Infractions	SE	p-value	Prop. same party	SE	p-value	Interaction	SE	p-value	N	Adj. R-squared
1	0.006	0.004	0.231	-0.119	0.031	0.031	0.023	0.011	0.134	5583	0.028
2	0.007	0.003	0.120	-0.066	0.041	0.210	-0.001	0.011	0.903	10939	0.025
3	0.005	0.001	0.019	-0.050	0.027	0.157	-0.009	0.005	0.146	13181	0.023
4	0.007	0.002	0.028	-0.024	0.034	0.540	-0.015	0.007	0.135	13630	0.024
5	0.006	0.002	0.036	-0.080	0.048	0.194	-0.009	0.008	0.304	13463	0.026
6	0.006	0.002	0.034	-0.104	0.052	0.141	-0.008	0.008	0.363	13111	0.027
7	0.006	0.002	0.029	-0.091	0.069	0.281	-0.010	0.006	0.186	12677	0.028
8	0.005	0.001	0.035	-0.146	0.098	0.233	-0.008	0.006	0.280	12173	0.029
9	0.005	0.001	0.037	-0.172	0.139	0.303	-0.007	0.007	0.342	11579	0.029
10	0.005	0.002	0.069	-0.213	0.245	0.448	-0.006	0.009	0.597	10926	0.030

Table B2: The effect of nearby corruption in interaction with the proportion of same-party audited neighbors.

	1. 2004	2. 2008	3. 2012	4. 2016
Intercept	0.08*	−0.00	0.00	0.04*
	(0.04)	(0.06)	(0.03)	(0.02)
Infractions	0.00	0.01*	0.00*	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)
Audited	0.19	0.12	0.03	0.03
	(0.14)	(0.15)	(0.09)	(0.09)
Interaction	−0.01	−0.01	−0.00	−0.00
	(0.01)	(0.01)	(0.00)	(0.00)
R^2	0.00	0.00	0.00	0.01
Adj. R^2	0.00	0.00	0.00	0.01
Num. obs.	3198	2841	2536	4180
RMSE	0.33	0.41	0.27	0.34

*$p < 0.05$

Table B3: The effect of nearby corruption at the ninth cumulative contiguity order on party switching by election year.

Contiguity	Estimate	SE	p-value	N	Adj. R-squared
1	0.0003777	0.0028322459	0.9023484	4800	0.015739728
2	-0.0039188	0.0018639599	0.1262748	9498	0.020978950
3	-0.0031956	0.0016857392	0.1542871	11448	0.023192008
4	-0.0033966	0.0013254411	0.0830252	11840	0.023957125
5	-0.0027960	0.0012922092	0.1191573	11688	0.023353135
6	-0.0022067	0.0009650581	0.1062899	11360	0.023381747
7	-0.0020373	0.0008133223	0.0873275	10962	0.023727303
8	-0.0018979	0.0007480084	0.0848838	10501	0.023662663
9	-0.0016069	0.0006857169	0.1009144	9960	0.023465109
10	-0.0014288	0.0006945120	0.1318476	9373	0.023839742

Table B4: The effect of nearby corruption on seeking reelection among incumbents who don't switch parties,

Contiguity	Infractions	SE	p-value	Party switch	SE	p-value	Interaction	SE	p-value	N	Adj. R-squared
1	-0.03	0.01	0.02	0.06	0.08	0.54	0.02	0.02	0.32	2221	0.08
2	-0.02	0.00	0.02	0.06	0.08	0.50	0.01	0.01	0.29	4387	0.08
3	-0.01	0.00	0.03	0.01	0.08	0.87	0.01	0.01	0.13	5332	0.09
4	-0.01	0.00	0.05	-0.00	0.08	0.97	0.01	0.01	0.12	5526	0.09
5	-0.01	0.00	0.07	-0.00	0.07	0.99	0.01	0.00	0.15	5467	0.09
6	-0.01	0.00	0.09	-0.01	0.07	0.91	0.01	0.00	0.16	5332	0.10
7	-0.01	0.00	0.12	0.01	0.07	0.90	0.01	0.00	0.23	5145	0.10
8	-0.01	0.00	0.14	-0.01	0.07	0.94	0.01	0.00	0.19	4933	0.10
9	-0.01	0.00	0.15	-0.00	0.07	0.95	0.00	0.00	0.20	4691	0.10
10	-0.01	0.00	0.23	-0.01	0.08	0.91	0.00	0.00	0.20	4428	0.10

Table B5: The effect of nearby corruption on securing reelection (conditional on seeking reelection) by whether the incumbent switches parties.

C. Descriptive statistics and robustness checks

Comparing audited and non-audited municipalities

Table C1 compares non-audited and audited municipalities across selected covariates. Most differences in means are indistinguishable from zero or negligible, with two exceptions. First, non-audited municipalities have, in average, about 24 more neighbors at the ninth cumulative contiguity order (the most generous definition of nearby implied by the adaptive lasso). However, they also have fewer audited neighbors and infractions per neighbors at that range, which only goes against the argument of the paper.

Second, non-audited municipalities have, in average, a higher population. This is mostly because the CGU anti-corruption program randomly selects municipalities with less than 500 thousand inhabitants. Similarly, the average differences in the proportion of non-audited and audited municipalities across election years are primarily a function of changes in the frequency of lotteries over time.

	Non-Audited	Audited	Difference	p-value
Neighbors	401.70	377.87	23.82	0.00
Audited neighbors	0.56	0.78	-0.22	0.00
Infractions per neighbor	1.22	1.59	-0.38	0.00
Population (thousands)	40.02	25.17	14.85	0.58
Female population	0.49	0.49	0.00	0.03
Rural population	0.38	0.38	0.00	1.00
Human Development Index	0.70	0.68	0.01	0.00
GDP per capita	13.42	11.64	1.78	0.00
Welfare benefits per capita	0.11	0.11	0.01	1.00
Share illiterate	0.22	0.24	-0.02	0.00
Share with college degree	0.03	0.03	0.00	0.00
Previous vote margin	0.15	0.15	0.00	1.00
PT incumbent	0.08	0.07	0.01	1.00
PSDB incumbent	0.16	0.14	0.02	1.00
Minas Gerais	0.16	0.12	0.04	0.00
São Paulo	0.13	0.10	0.02	0.57
Northeast	0.36	0.42	-0.06	0.00
2004 election	0.28	0.18	0.10	0.00
2008 election	0.27	0.50	-0.23	0.00
2012 election	0.19	0.20	-0.02	1.00
2016 election	0.27	0.12	0.14	0.00

Table C1: Comparing non-audited and audited municipalities across selected covariates. P-values adjusted using the Bonferroni correction

Logistic regression

This section shows that main results are robust to using logistic regression. The error correction and fixed-effects schemes remain the same. The list below details the correspondence with the main text:

- Table C2 corresponds to Figure 2 in the main text and Table B1 in this appendix.
- Table C3 corresponds to Figure 3 in the main text and Table B2 in this appendix.
- Table C4 corresponds to Figure 4 in the main text and Table B3 in this appendix.
- Tables C5 and C6 correspond to Figure 5 in the main text and Tables B4 and B5 in this appendix.

Contiguity	Infractions	SE	p-value	Audited	SE	p-value	Interaction	SE	p-value
1	0.085	0.038	0.024	0.192	0.400	0.631	-0.047	0.038	0.024
2	0.065	0.006	0.000	0.238	0.080	0.003	-0.049	0.006	0.000
3	0.041	0.008	0.000	0.031	0.145	0.830	-0.007	0.008	0.000
4	0.050	0.008	0.000	0.084	0.142	0.552	-0.012	0.008	0.000
5	0.049	0.006	0.000	0.011	0.157	0.946	-0.004	0.006	0.000
6	0.049	0.006	0.000	0.161	0.123	0.192	-0.013	0.006	0.000
7	0.046	0.005	0.000	0.202	0.121	0.094	-0.014	0.005	0.000
8	0.042	0.005	0.000	0.338	0.058	0.000	-0.018	0.005	0.000
9	0.039	0.004	0.000	0.283	0.120	0.019	-0.014	0.004	0.000
10	0.036	0.004	0.000	0.358	0.177	0.044	-0.015	0.004	0.000

Table C2: Reproducing results from Figure 2 and Table B1 using logistic regression.

Contiguity	Infractions	SE	p-value	Audited	SE	p-value	Interaction	SE	p-value
1	0.060	0.036	0.089	-1.284	0.431	0.003	0.261	0.036	0.089
2	0.062	0.019	0.001	-0.932	0.625	0.136	0.039	0.019	0.001
3	0.047	0.002	0.000	-0.852	0.464	0.067	-0.042	0.002	0.000
4	0.062	0.004	0.000	-0.673	0.497	0.176	-0.096	0.004	0.000
5	0.051	0.008	0.000	-1.796	0.610	0.003	-0.008	0.008	0.000
6	0.048	0.007	0.000	-2.510	0.366	0.000	0.021	0.007	0.000
7	0.046	0.007	0.000	-2.643	0.736	0.000	0.002	0.007	0.000
8	0.037	0.008	0.000	-4.122	0.905	0.000	0.054	0.008	0.000
9	0.033	0.008	0.000	-4.755	1.476	0.001	0.063	0.008	0.000
10	0.029	0.012	0.017	-5.566	2.845	0.050	0.083	0.012	0.017

Table C3: Reproducing results from Figure 3 and Table B2 using logistic regression.

	1. 2004	2. 2008	3. 2012	4. 2016
Intercept	−2.37*	−2.56*	−3.54*	−2.72*
	(0.35)	(0.36)	(0.40)	(0.17)
Infractions	0.02	0.05*	0.04*	0.04*
	(0.02)	(0.01)	(0.01)	(0.01)
Audited	1.99	0.72	0.54	0.21
	(1.38)	(0.87)	(1.11)	(0.90)
Interaction	−0.11	−0.03	−0.02	−0.01
	(0.06)	(0.04)	(0.04)	(0.04)
AIC	2428.83	2973.26	1414.39	3328.56
BIC	2453.11	2997.07	1437.74	3353.91
Log Likelihood	-1210.42	-1482.63	-703.20	-1660.28
Deviance	2420.83	2965.26	1406.39	3320.56
Num. obs.	3198	2841	2536	4180

*$p < 0.05$

Table C4: Reproducing results from Figure 4 and Table B3 using logistic regression.

Contiguity	Infractions	SE	p-value
1	0.002	0.013	0.894
2	-0.018	0.009	0.033
3	-0.015	0.008	0.061
4	-0.016	0.006	0.014
5	-0.013	0.006	0.042
6	-0.010	0.005	0.034
7	-0.009	0.004	0.023
8	-0.009	0.004	0.020
9	-0.007	0.003	0.038
10	-0.006	0.004	0.066

Table C5: Reproducing results from panel A of Figure 5 and Table B4 using logistic regression.

Contiguity	Infractions	SE	p-value	Party switch	SE	p-value	Interaction	SE	p-value
1	-0.129	0.029	0.000	0.473	0.645	0.463	0.091	0.104	0.380
2	-0.091	0.007	0.000	0.308	0.485	0.526	0.082	0.070	0.237
3	-0.080	0.016	0.000	0.083	0.470	0.860	0.077	0.045	0.087
4	-0.065	0.018	0.000	0.007	0.490	0.988	0.066	0.037	0.079
5	-0.052	0.018	0.005	0.035	0.484	0.942	0.049	0.033	0.133
6	-0.047	0.020	0.016	0.011	0.505	0.983	0.042	0.029	0.147
7	-0.036	0.018	0.047	0.128	0.538	0.812	0.028	0.025	0.257
8	-0.035	0.019	0.059	0.081	0.573	0.887	0.028	0.023	0.215
9	-0.030	0.017	0.072	0.119	0.594	0.841	0.023	0.020	0.243
10	-0.027	0.019	0.162	0.154	0.683	0.822	0.020	0.018	0.282

Table C6: Reproducing results from panel B of Figure 5 and Table B5 using logistic regression.

D. Defining nearby

Section 4.2 uses cumulative orders of contiguity to show that effects are robust to different definitions of exposure to nearby corruption up to the tenth level of contiguity. This does not imply that non-audited municipalities are affected by revealing corruption among neighbors that are 10 municipalities apart. Rather, it suggests that the cumulative patterns are consistent up to that point.

I approach the question of how far spillovers travel as a variable selection problem in the context of supervised statistical learning (more about this in future work, stay tuned!). The task is to determine how many orders of contiguity should the true model include. Of course, the true model is not observed, but we can compare alternative specifications and choose the model that best fits the data.

Figure 1 in the main text reports shows that the distribution of nearby corruption is similar across the contiguity orders consider in this protocol. For completeness. Figure D1 shows the distribution of total and audited neighbors in those ranges. The median number of total neighbors at the first contiguity order is 6, while at the tenth is 85. The median number of audited neighbors is 1 and 8, respectively. While the increase in the total number of neighbors conveys how far away the tenth order of contiguity reaches, the modest increase in the distribution of audited neighbors suggests that results are not an artifice of increasing the number of audited neighbors.

I fit an adaptive lasso (Zou 2006) in the subset of non-audited municipalities with an indicator of whether the incumbent seeks reelection under a different party as the outcome variable. The predictors record the number of corrupt infractions per audited neighbor in a given contiguity order, so that the first predictor includes infractions among the audited neighbors with whom a municipality shares borders. The second predictor includes infractions among the neighbors of the immediate neighbors, and so on, up to the tenth order of contiguity (Figure D1 suggests this is very far away). Note that these variables are different from the one used in the paper, as they are not cumulative, but rather contain infractions only within that contiguity order.

I include OLS regression coefficients as weights for the adaptive lasso. While different sets of weights or constraints are possible, for example, allowing a coefficient for a given contiguity order to be non-zero only if the coefficient for the previous contiguity order is non-zero as well. I opt instead for an agnostic approach that lets the data speak.

Figure D2 shows the adaptive lasso path along increasing values of the shrinkage parameter λ. As λ increases, coefficients shrink towards zero. I choose the values of λ that minimize the root mean squared error (RMSE) via 10-fold cross-validation. The vertical dashed lines denote the value of λ that minimizes RMSE (left) and

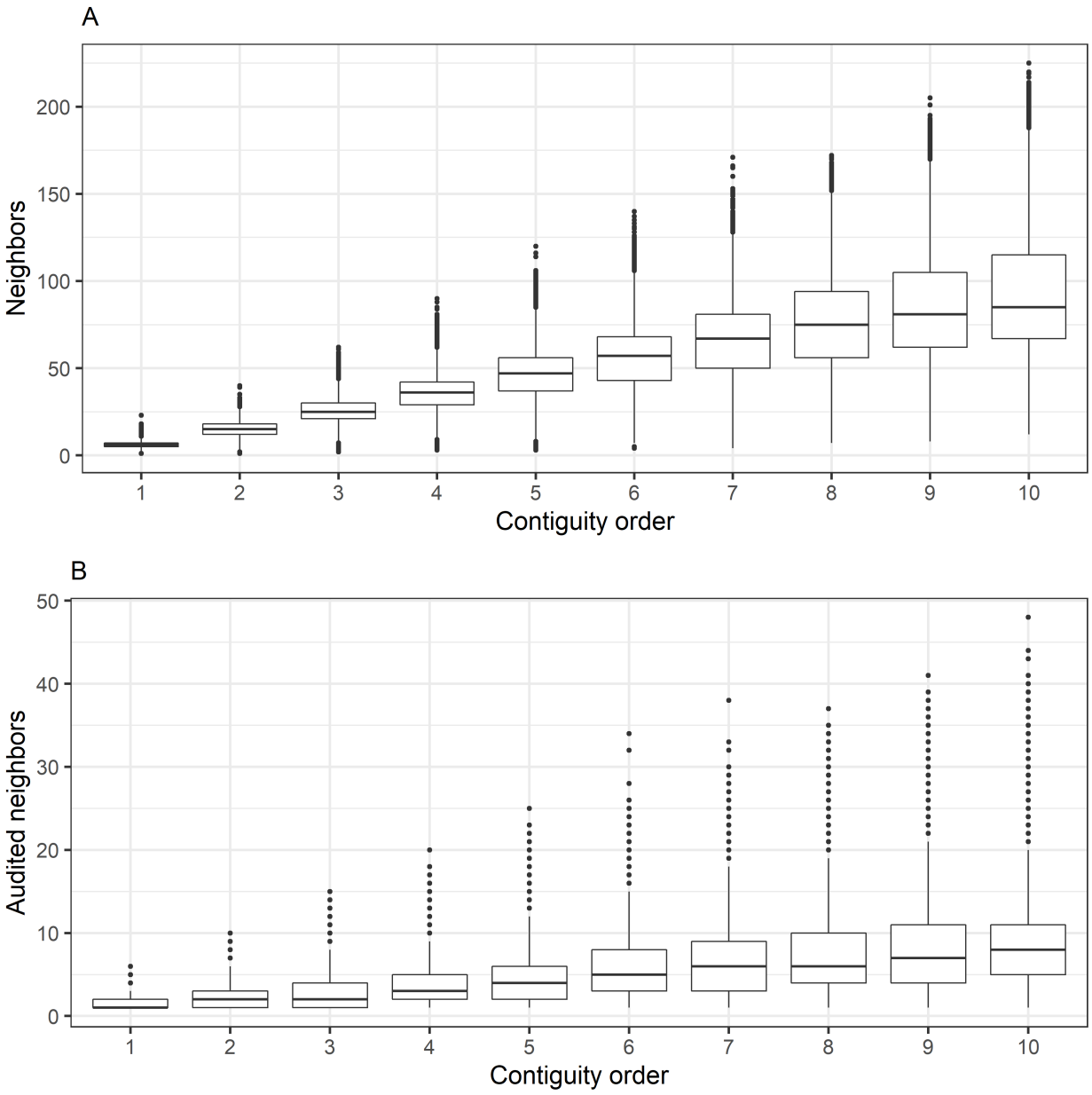

Figure D1: Distribution of total number of neighbors (panel A) and number of audited neighbors (panel B) by contiguity order.

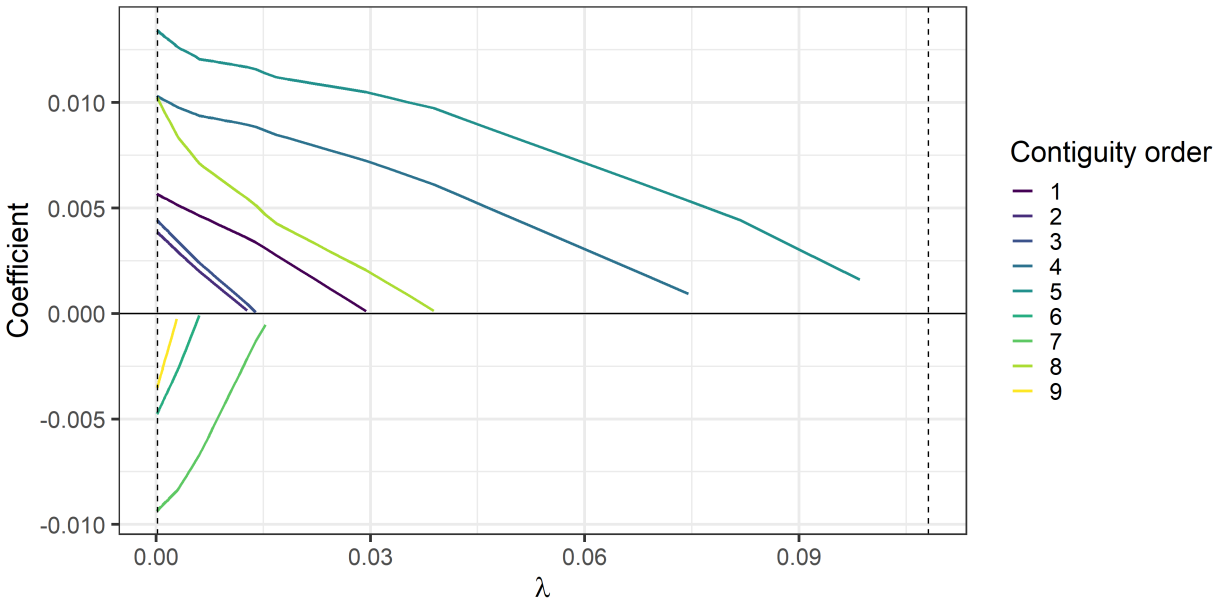

Figure D2: Lasso path fit for the incumbent party seeking reelection as the outcome and number of infractions per audited neighbor.

the largest λ within one standard deviation from it.

Any non-zero coefficients within this range are plausible candidates for the true model specification. The standard practice is to settle on the largest value of λ within a standard deviation from the smallest RMSE to avoid overfitting. At this value of lambda, only the fifth order of contiguity come close to being non-zero. Therefore, the narrowest possible definition of nearby within the optimal λ range considers cumulative nearby corruption up to the fifth order.

Within one standard deviation λ, all predictors except for nearby corruption at the tenth contiguity order are non-zero. Therefore, the most generous model within the optimal λ range considers cumulative nearby corruption up to the ninth contiguity order.

References

Rundlett, Ashlea P. 2018. "The Effects of Revealed Corruption on Voter Attitudes and Participation: Evidence from Brazil." Ph.D. Dissertation, University of Illinois at Urbana-Champaign. http://hdl.handle.net/2142/101330.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101 (476): 1418–29.