Building a Robot Judge: Data Science for Decision-Making

12. Algorithms and Decisions III

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.
- a subjective label, such as "harmful to self or others", when made by a human, could be biased (and so would teaching an ML model to reproduce that label)

Data can be biased

- Education:
  - SAT scores might be used to guide college admissions, but some students get SAT prep courses
  - Teachers (grading essays) might be biased against some students $\rightarrow$ so will automated essay graders based on those grades.
- Criminal risk scoring (Skeem and Lovenkamp 2016):
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky due to longer criminal histories (**which were produced by biased system**).
  - similarly: we measure recidivism as "is re-arrested" rather than "commits more crimes". some people more likely to be re-arrested due to policing bias.
  - selective labeling:
    - predictive policing – produces evidence of more crimes in the neighborhoods where police want to go.
    - only observe recidivism if released.
- a subjective label, such as "harmful to self or others", when made by a human, could be biased (and so would teaching an ML model to reproduce that label)

These types of problems cannot be fixed by ML.
But ML can help diagnose them.

Discussion

- Algorithms can help us understand if human judges make mistakes, and diagnose reasons for bias.

Discussion

- Algorithms can help us understand if human judges make mistakes, and diagnose reasons for bias.
- Not just about prediction. Key is starting with decision:
  - Performance benchmark: Current "human" decisions

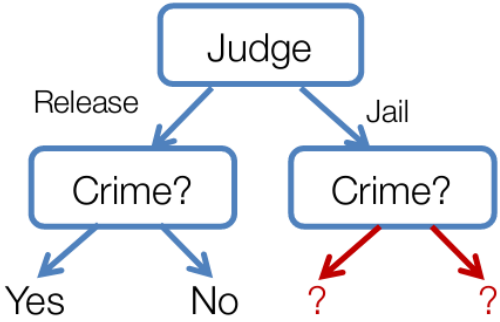
Discussion

- Algorithms can help us understand if human judges make mistakes, and diagnose reasons for bias.
- Not just about prediction. Key is starting with decision:
  - Performance benchmark: Current "human" decisions
- Question: What are we really optimizing?

Focusing on re-arrest rates is limited

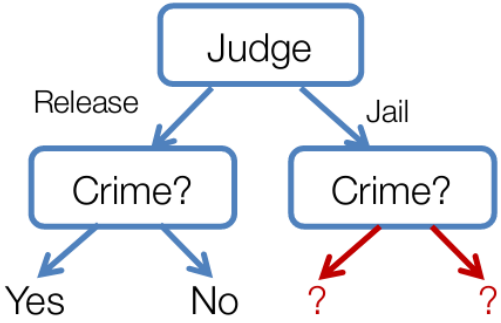
- Is minimizing the crime rate really the right goal?
- There are other important factors
  - Consequences of jailing on the family
  - Jobs and the workplace
  - Future behavior of the defendant
- How could we measure/model these?

Problem: Judge is selectively labeling the dataset



- We can only train on released people:
  - By jailing, judge is selectively hiding labels!

Problem: Judge is selectively labeling the dataset

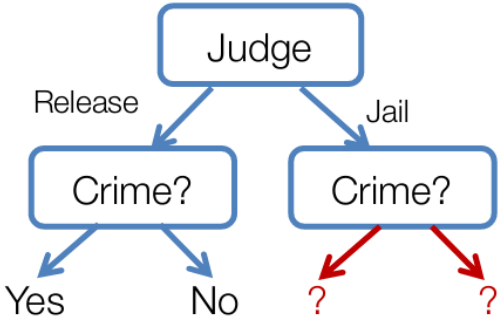


- We can only train on released people:
  - By jailing, judge is selectively hiding labels!

Selective labels introduce bias. Example:

- Say young people with no tattoos have no risk for crime. Judge releases them.
- Machine observes age, but does not observe tattoos.

Problem: Judge is selectively labeling the dataset



- We can only train on released people:
  - By jailing, judge is selectively hiding labels!

Selective labels introduce bias. Example:

- Say young people with no tattoos have no risk for crime. Judge releases them.
- Machine observes age, but does not observe tattoos.
- Machine would falsely conclude that all young people do no crime, and release all young people.

Solution: Contraction

- Selection problem is one-sided: We observe counterfactual (crime rate) for released defendants, but not jailed defendants.

Solution: Contraction

- Selection problem is one-sided: We observe counterfactual (crime rate) for released defendants, but not jailed defendants.

- **Contraction**:
  - Take released population of a lenient judge.
  - Then ask which additional defendant we would jail to minimize crime rate.
  - Compare change in crime rate to that observed for stricter judge.
- **Why does this approach require random assignment of cases to judges to work?**

Comparing Machine Judges (Left Panel) to Human Judges (Right Panel)

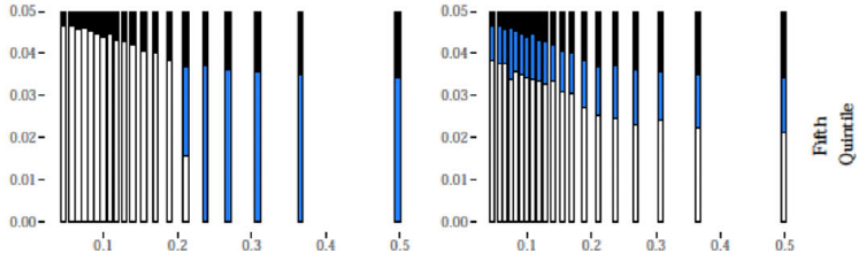


FIGURE VI

Who Do Stricter Judges Jail and Who Would the Algorithm Jail? Comparing Predicted Risk Distributions across Leniency Quintiles

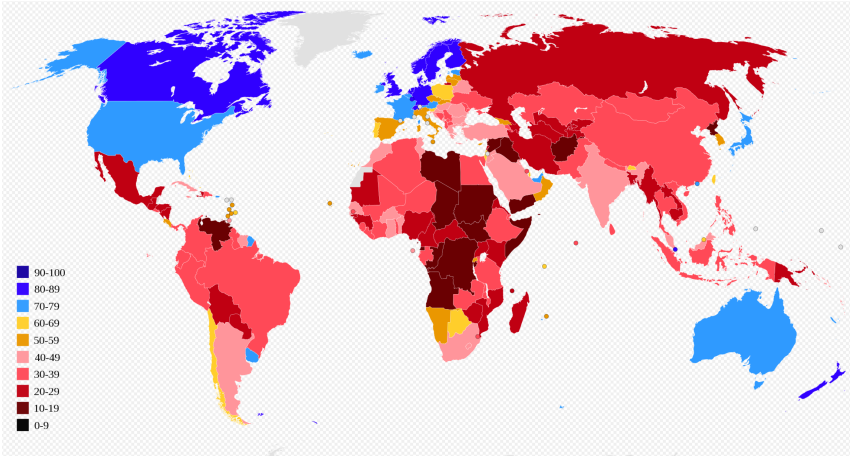
- black = even most lenient judges (bottom quintile) would jail this defendant.
- blue = additional jailed by the strictest judges (top quintile). left panel = algorithm, right panel = human judges.
- white = who is released by all judges

Labels are Driven by Decisions

- We don't see labels of people that are jailed
- This is a broader problem in policymaking systems:
  - Prediction $\rightarrow$ Decision $\rightarrow$ Outcome
- Which outcomes we see depends on our decisions

Outline

Motivation (Ash, Galletta, Giommoni 2020)



Corruption Perceptions Index, 2018

Global costs of corruption were $2.6 trillion in 2018, according to U.N. data.
Firms and individuals spend more than $1 trillion in bribes every year.

This Paper's Goals

- **Objective 1:** Predict fiscal corruption based on public finance accounts.
  - In Brazilian municipalities, we have information on fiscal corruption from random audits.
  - We train a machine learning algorithm to detect corruption in held-out data using budget data.

This Paper's Goals

- **Objective 1:** Predict fiscal corruption based on public finance accounts.
  - In Brazilian municipalities, we have information on fiscal corruption from random audits.
  - We train a machine learning algorithm to detect corruption in held-out data using budget data.
- **Objective 2:** Construct new measure of corruption for all municipalities and years (not just those that have been audited) and use for empirical analysis.
  - Effect of public transfers on corruption (IV).
  - Effect of audits on corruption (DD).

This Paper's Goals

- **Objective 1:** Predict fiscal corruption based on public finance accounts.
  - In Brazilian municipalities, we have information on fiscal corruption from random audits.
  - We train a machine learning algorithm to detect corruption in held-out data using budget data.
- **Objective 2:** Construct new measure of corruption for all municipalities and years (not just those that have been audited) and use for empirical analysis.
  - Effect of public transfers on corruption (IV).
  - Effect of audits on corruption (DD).
- **Objective 3:** Use predictions to analyze counterfactual audit policies.
  - What can be accomplished by targeting audits to municipalities with high-risk budgets?

Brazilian municipalities

- In Brazil, local municipalities ($N = 5563$) play a central role in government services:
  - e.g., primary education, healthcare, housing, transportation.

Brazilian municipalities

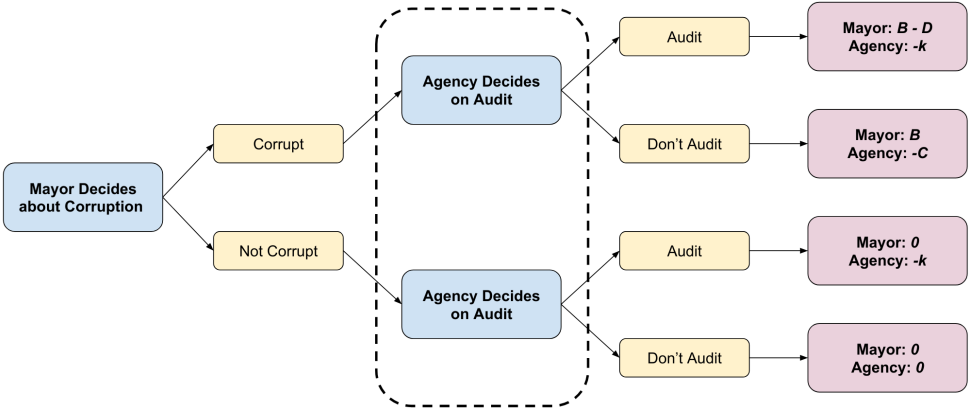
- In Brazil, local municipalities ($N = 5563$) play a central role in government services:
  - e.g., primary education, healthcare, housing, transportation.
- In 2003, Brazilian government introduced innovative anti-corruption program:
  - **Audit of public spending** in **randomly selected municipalities** (through public lottery).
  - team of 10-15 auditors spend two weeks in municipal offices.
  - they write a report, send to authorities for criminal penalties and make it public.

Outline

- Stage 1: Mayor decides whether to engage in corruption.
  - if corrupt, mayor gets payoff B, society loses C (zero otherwise).

- Stage 1: Mayor decides whether to engage in corruption.
  - if corrupt, mayor gets payoff B, society loses C (zero otherwise).
- Stage 2: Agency decides whether to audit municipality i.
  - if audit, agency pays cost k, zero otherwise
  - if audit reveals corruption:
    - society does not lose C; mayor pays penalty $D > B$

- Stage 1: Mayor decides whether to engage in corruption.
  - if corrupt, mayor gets payoff B, society loses C (zero otherwise).
- Stage 2: Agency decides whether to audit municipality i.
  - if audit, agency pays cost k, zero otherwise
  - if audit reveals corruption:
    - society does not lose C; mayor pays penalty $D > B$



- In game theory, this is called an "inspection game".

Matrix Form (chalk board)

Matrix Form (chalk board)

- There is no pure-strategy Nash equilibrium (cycling).

Matrix Form (chalk board)

- There is no pure-strategy Nash equilibrium (cycling).
- Assume mixed strategies:
  - p= probability of corruption, q= probability of audit.
  - **Mixed strategy equilibrium**: (p^*, q^*) such that each player is indifferent between options.

Matrix Form (chalk board)

- There is no pure-strategy Nash equilibrium (cycling).
- Assume mixed strategies:
  - p= probability of corruption, q= probability of audit.
  - **Mixed strategy equilibrium**: (p^*, q^*) such that each player is indifferent between options.
- Payoffs for mayor:
  - no corruption: 0
  - corruption: $\underbrace{q(B-D)}_{\text{audit}} + \underbrace{(1-q)B}_{\text{no audit}} = B - qD$

Matrix Form (chalk board)

- There is no pure-strategy Nash equilibrium (cycling).
- Assume mixed strategies:
  - p= probability of corruption, q= probability of audit.
  - **Mixed strategy equilibrium**: (p^*, q^*) such that each player is indifferent between options.
- Payoffs for mayor:
  - no corruption: 0
  - corruption: $\underbrace{q(B-D)}_{\text{audit}} + \underbrace{(1-q)B}_{\text{no audit}} = B - qD$

  $\rightarrow$ **equilibrium audit probability** $q^* = \frac{B}{D}$

Matrix Form (chalk board)

- There is no pure-strategy Nash equilibrium (cycling).
- Assume mixed strategies:
  - p= probability of corruption, q= probability of audit.
  - **Mixed strategy equilibrium**: (p^*, q^*) such that each player is indifferent between options.
- Payoffs for mayor:
  - no corruption: 0
  - corruption: $\underbrace{q(B-D)}_{\text{audit}} + \underbrace{(1-q)B}_{\text{no audit}} = B - qD$

  $\rightarrow$ **equilibrium audit probability** $q^* = \frac{B}{D}$
- Similarly, payoffs for agency:
  - audit: $p(-k) + (1-p)(-k) = -k$
  - no audit: $p(-C)$

Matrix Form (chalk board)

- There is no pure-strategy Nash equilibrium (cycling).
- Assume mixed strategies:
  - p= probability of corruption, q= probability of audit.
  - **Mixed strategy equilibrium**: (p^*, q^*) such that each player is indifferent between options.
- Payoffs for mayor:
  - no corruption: 0
  - corruption: $\underbrace{q(B-D)}_{\text{audit}} + \underbrace{(1-q)B}_{\text{no audit}} = B - qD$

  $\rightarrow$ **equilibrium audit probability** $q^* = \frac{B}{D}$
- Similarly, payoffs for agency:
  - audit: $p(-k) + (1-p)(-k) = -k$
  - no audit: $p(-C)$

  $\rightarrow$ **equilibrium corruption probability** $p^* = \frac{k}{C}$

Equilibrium Audit Policy

- Equilibrum of game:
  - **corruption probability** $p^* = \frac{k}{C}$
  - **audit probability** $q^* = \frac{D}{B}$

$\rightarrow$ Randomly assigned audits to a fraction q^* of municipalities is the equilibrium audit policy.

Equilibrium Audit Policy

- Equilibrum of game:
  - **corruption probability** $p^* = \frac{k}{C}$
  - **audit probability** $q^* = \frac{D}{B}$

$\rightarrow$ Randomly assigned audits to a fraction q^* of municipalities is the equilibrium audit policy.

- Note that the observed corruption rate is

$$p^* = \frac{1}{N} \sum_{i=1}^{N} p_i$$

  the average of p_i, the probability of corruption for municipality i.
- Below, we will consider how this changes if agency can guess $\hat{p}(X_i)$ based on budget factors X_i.

Outline

Corruption Audit Data

- Municipal audit reports are available from the agency web site:

	DOWNLOAD	TÍTULO	LINHA DE ATUAÇÃO	PUBLICADO EM	MUNICÍPIOS	TRECHOS
☐		Relatório de Fiscalização Sorteio de Municípios - Olho D'Água das Flores/AL	Fiscalização em Entes Federativos - Municípios	18/12/2018	OLHO D'ÁGUA DAS FLORES - AL	Assim como ocorre em muitos **municípios**, Olho d'Água das
☐		Relatório de fiscalização nº 201800789 - General Maynard/SE	Fiscalização em Entes Federativos - Municípios	05/12/2018	GENERAL MAYNARD - SE	Junho **de** 2007, são os **municípios** os responsáveis pelo cadastramento
☐		Relatório de Fiscalização Sorteio de Municípios - Vargem Alegre/MG	Fiscalização em Entes Federativos - Municípios	17/10/2017	VARGEM ALEGRE - MG	**de** outros **municípios**.

- Brollo et al (2013) construct corruption labels from the reports for 1481 audited municipalities, 2003-2010. Their data is online.

Local Budget Data

- The annual municipality budget is available from various web sites:
  - We collected/cleaned data for 2001 through 2012 and made them comparable across years.

Local Budget Data

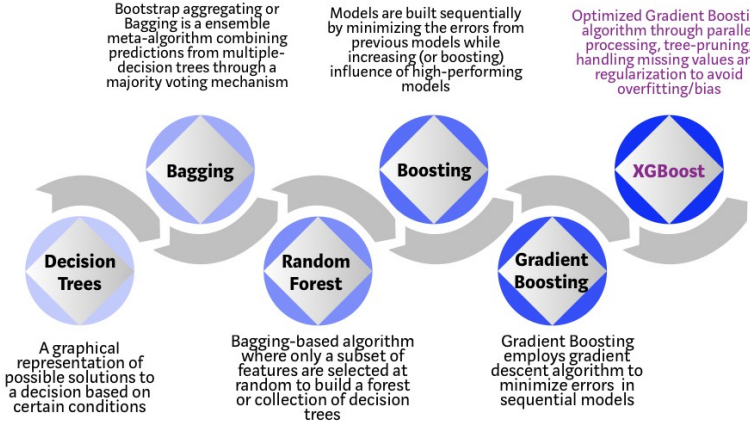
- The annual municipality budget is available from various web sites:
  - We collected/cleaned data for 2001 through 2012 and made them comparable across years.
- In total we have 797 budget variables:
  - Revenue 250, Expenditure 334, Active 100, Passive 79.

Gradient Boosted Classifier

- Gradient boosting classifier (GBC): ensemble of decision trees (Friedman, 2001; Hastie et al 2009).
  - same model used by Kleinberg et al (QJE 2018) to predict criminal recidivism.

Gradient Boosted Classifier

- Gradient boosting classifier (GBC): ensemble of decision trees (Friedman, 2001; Hastie et al 2009).
  - same model used by Kleinberg et al (QJE 2018) to predict criminal recidivism.
- We use XGBoost ("Extreme Gradient Boosting"), an optimized python implementation (Chen and Guestrin 2016).
  - Feurer et al (2018) find that XGBoost beats a sophisticated AutoML procedure with grid search over 15 classifiers and 18 data preprocessors.



Complicated in theory, easy in practice

```
from xgboost import XGBClassifier
model = XGBClassifier()

model.fit(X_train, y_train,
          early_stopping_rounds=10,
          eval_metric="logloss",
          eval_set=[(X_eval, y_eval)]
          )

y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
```

Model Training

1. Shuffle dataset into 80% training set and 20% test set
   - budget predictors standardized to mean zero and variance one in training set
2. Tuned hyperparameters in the training set using five-fold cross-validation (e.g., max depth of trees and learning rate)
   - Use early stopping to avoid over-fitting.
3. Take tuned model and get performance metrics in the test set

Model Performance in Test Set

	Guess "Not Corrupt"	*OLS*	XGBoost
Accuracy	0.58	0.594	0.750
AUC-ROC	0.5		
F1	0.0		

- Test-set accuracy of $\sim 75\%$ is much better than guessing (58%) or predictions from OLS (59%)

Model Performance in Test Set

	Guess "Not Corrupt"	*OLS*	XGBoost
Accuracy	0.58	0.594	0.750
AUC-ROC	0.5	0.562	**0.814**
F1	0.0	0.413	0.665

- AUC-ROC ("Area under the receiver operating curve") is a standard metric, ranging from 0.5 (guessing) ato 1.0 (perfect accuracy).
  - Interpretation: probability that a randomly sampled corrupt municipality is ranked more highly by predicted probability of corruption than a randomly sampled non-corrupt municipality.
  - **AUC$\approx$.81 is better than Kleinberg et al (QJE 2018) who report AUC=0.707.**

Confidence Intervals on ML Metrics

- nested cross-validation with 5 folds $\rightarrow$ produce 5 sets of perfomance metrics.

Confidence Intervals on ML Metrics

- nested cross-validation with 5 folds $\rightarrow$ produce 5 sets of perfomance metrics.

Metric	**Accuracy**	**AUC**
Mean	0.74	0.81
Median	0.74	0.82
S.D. / S.E.	0.01	0.02
95% CI's	[.73 .75]	[.79 .83]

Confidence intervals constructed as mean +/- 2$\times$S.E..

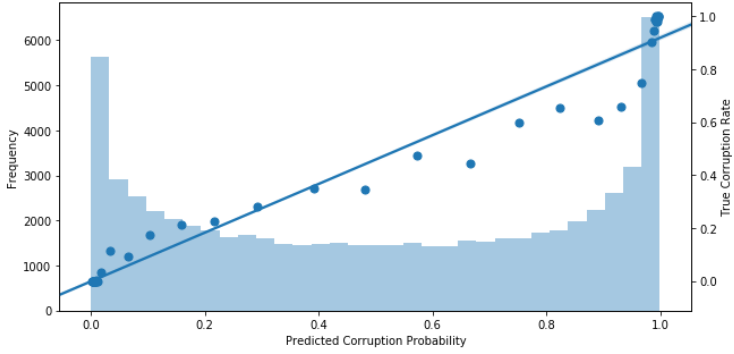
Confusion Matrix for Test-Set Predictions

	Prediction	
Truth	Not Corrupt	Corrupt
Not Corrupt	**614**	100
Corrupt	185	**313**

Confusion Matrix for Test-Set Predictions

	Prediction	
Truth	Not Corrupt	Corrupt
Not Corrupt	**614**	100
Corrupt	185	**313**

True Corrupt Rate vs Predicted Prob. Corruption



Most Important Budget Features for Corruption Prediction

(Ash, Galletta, Giommoni 2020)

Most Important Budget Features for Corruption Prediction

(Ash, Galletta, Giommoni 2020)

Category	Macro Category	Weight	Category	Macro Category	Weight
Assets	Assets	330	Outstanding loan credit	Assets	69.4
Financial assets	Assets	182	Tax on industrialized products	Revenue	69
Population		142.6	**Property tax on land/buildings**	Revenue	68
Cash	Assets	116.4	Liquid assets	Assets	67.8
Spending in agriculture	Expenditure	94.8	Civil servant per diems	Expenditure	67.4
Property tax on rural land	Revenue	89.6	Spending for legislative procedure	Expenditure	65
Bank deposit	Assets	85.4	Taxes	Revenue	64.4
Motor vehicle property tax (from FG)	Revenue	72.8	Budget deficit		63
Transf. of ownership tax	Revenue	72	Non financial current asset	Assets	60.6
Spending in transportation	Expenditure	72	**Capital expenditure**	Expenditure	60

Important Features tend to show up in Audit Reports

We scraped all of the municipal audit reports from the agency web site.

- After converting the PDFs to text and some mild pre-processing, we counted the mentions of different budget features in the reports.

Important Features tend to show up in Audit Reports

We scraped all of the municipal audit reports from the agency web site.

- After converting the PDFs to text and some mild pre-processing, we counted the mentions of different budget features in the reports.

Produce dataset:
{budget feature, audit report mentions, feature importance}

- Regress **audit report mentions** against **XGBoost feature importance.**

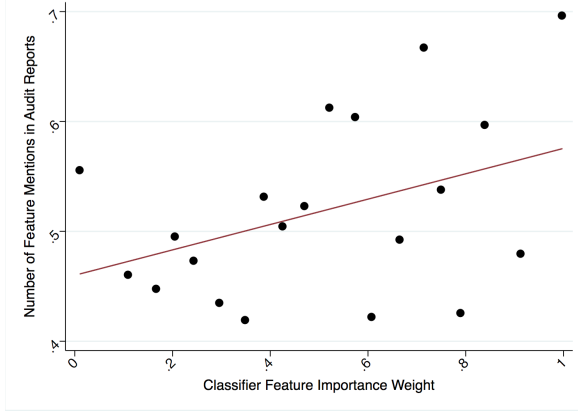
Important Features tend to show up in Audit Reports

We scraped all of the municipal audit reports from the agency web site.

- After converting the PDFs to text and some mild pre-processing, we counted the mentions of different budget features in the reports.

Produce dataset:
{budget feature, audit report mentions, feature importance}

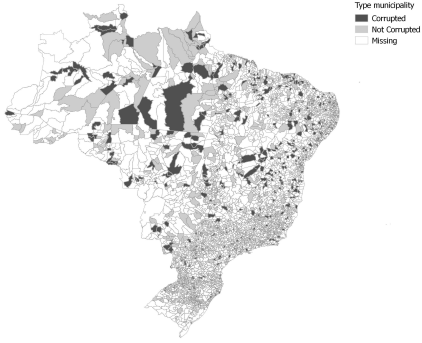
- Regress **audit report mentions** against **XGBoost feature importance.**



Notes: Binscatter for frequency that budget feature appears in the municipal audit reports (vertical axis) against binned feature importance weights for each feature (horizontal axis). Pearson's correlation is 0.17 (.24 for the log measures, rather than ranks). Slope coefficient is 0.112 with p=.03 (robust standard errors).

Outline

Applying to Full Dataset



(a) Actual Corruption

(b) Predicted Corruption

We regressed predicted corruption in pre-audit years on having an audit, and there was no difference in any specification (consistent with randomization of audits).

Analysis 1: Revenue Shocks and Corruption

- Brollo et al (2013) find that a **windfall of public revenues** (federal transfers) leads to an increase in rent-seeking by the public administration (*i.e.* subsequent increase in corruption).

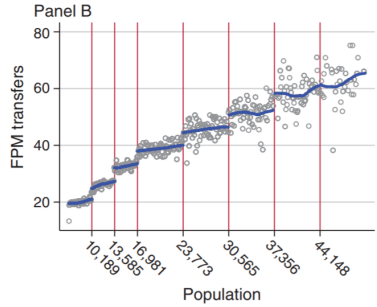
Analysis 1: Revenue Shocks and Corruption

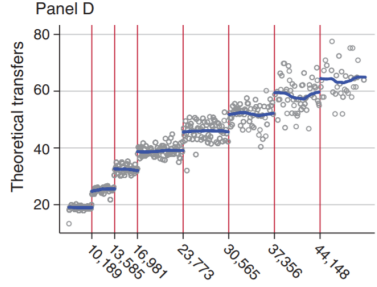
- Brollo et al (2013) find that a **windfall of public revenues** (federal transfers) leads to an increase in rent-seeking by the public administration (*i.e.* subsequent increase in corruption).

- Empirical Strategy: Fuzzy RDD
  - Exogenous variation in transfers due to discrete population thresholds.
  - imperfect takeup, so instrument actual transfers τ_i with prescribed transfers z_i

Analysis 1: Revenue Shocks and Corruption

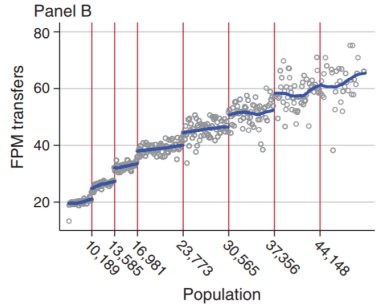
- Brollo et al (2013) find that a **windfall of public revenues** (federal transfers) leads to an increase in rent-seeking by the public administration (*i.e.* subsequent increase in corruption).
- Empirical Strategy: Fuzzy RDD
  - Exogenous variation in transfers due to discrete population thresholds.
  - imperfect takeup, so instrument actual transfers τ_i with prescribed transfers z_i

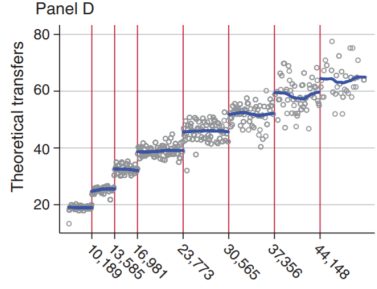




Analysis 1: Revenue Shocks and Corruption

- Brollo et al (2013) find that a **windfall of public revenues** (federal transfers) leads to an increase in rent-seeking by the public administration (*i.e.* subsequent increase in corruption).
- Empirical Strategy: Fuzzy RDD
  - Exogenous variation in transfers due to discrete population thresholds.
  - imperfect takeup, so instrument actual transfers τ_i with prescribed transfers z_i





- **Our extension**: Analyze **universe** of Brazilian municipalities (not only those being audited). N increases from 1115 to 5563.

Fuzzy RD (IV) Estimating Equations

- First stage: impact of prescribed transfers (z_i) on actual transfers (τ_i)

$$\tau_i = g(P_i) + \gamma z_i + u_i \tag{1}$$

- Second stage: impact of instrumented actual transfers (τ_i) on ML-predicted corruption (y_i)

$$y_i = g(P_i) + \beta \tau_i + \epsilon_i \tag{2}$$

  – polynomial $g(\cdot)$ in population P_i

Activity: Exogeneity/Exclusion

https://padlet.com/eash44/cfsa9e4m4lycv33f

$$\tau_i = g(P_i) + \gamma z_i + u_i$$
$$y_i = g(P_i) + \beta\tau_i + \epsilon_i$$

- Last Name starts with A-M:
  - Articulate exogeneity assumption, and a potential violation.
- Last Name starts with N-Z:
  - Articulate exclusion restriction, and a potential violation.

Brollo et al (2013) Replication: First Stage

	Audited cities (1)	All cities (2)	Never Audited (3)
Panel A. First Stage			
Prescribed transfers	0.680***	0.687***	0.700***
	(0.021)	(0.022)	(0.023)
Observations	1115	5563	4693

Standard errors clustered at the municipal level are in parentheses: *p<0.10, ** p<0.05, *** p<0.01. Prescribed transfers (z_i), actual transfers (τ_i), predicted corruption (y_i). First stage: $\tau_i = g(P_i) + \alpha_\tau z_i + \delta_t + \gamma_s + u_i$; Second stage: $y_i = g(P_i) + \beta_y \tau_i + \delta_t + \gamma_s + \epsilon_i$; polynomial $g(\cdot)$ in population P_i, time fixed effects δ_t, state fixed effects γ_s (as in Brollo et al. 2013).

Brollo et al (2013) Replication: Audited Cities

	Audited cities (1)	All cities (2)	Never Audited (3)
Panel A. First Stage			
Prescribed transfers	0.680***	0.687***	0.700***
	(0.021)	(0.022)	(0.023)
Panel B. Reduced Form			
Prescribed transfers	**0.00526****		
	(0.00264)		
Panel C. 2SLS			
Actual transfers	**0.00862****		
	(0.004)		
Observations	1115	5563	4693

Standard errors clustered at the municipal level are in parentheses: *p<0.10, ** p<0.05, *** p<0.01. Prescribed transfers (z_i), actual transfers (τ_i), predicted corruption (y_i). First stage: $\tau_i = g(P_i) + \alpha_\tau z_i + \delta_t + \gamma_s + u_i$; Second stage: $y_i = g(P_i) + \beta_y \tau_i + \delta_t + \gamma_s + \epsilon_i$; polynomial $g(\cdot)$ in population P_i, time fixed effects δ_t, state fixed effects γ_s (as in Brollo et al. 2013).

Brollo et al (2013) Replication: Never-Audited Cities

	Audited cities (1)	All cities (2)	**Never Audited** (3)
Panel A. First Stage			
Prescribed transfers	0.680***	0.687***	0.700***
	(0.021)	(0.022)	(0.023)
Panel B. Reduced Form			
Prescribed transfers	0.00526**	0.00370***	**0.00294*** **
	(0.00264)	(0.001)	**(0.001)**
Panel C. 2SLS			
Actual transfers	0.00862**	0.00731***	**0.00660*** **
	(0.004)	(0.001)	**(0.001)**
Observations	1115	5563	4693

Standard errors clustered at the municipal level are in parentheses: *p<0.10, ** p<0.05, *** p<0.01. Prescribed transfers (z_i), actual transfers (τ_i), predicted corruption (y_i). First stage: $\tau_i = g(P_i) + \alpha_\tau z_i + \delta_t + \gamma_s + u_i$; Second stage: $y_i = g(P_i) + \beta_y \tau_i + \delta_t + \gamma_s + \epsilon_i$; polynomial $g(\cdot)$ in population P_i, time fixed effects δ_t, state fixed effects γ_s (as in Brollo et al. 2013).

Analysis 2: Effects of auditing on subsequent corruption

- ML-predicted corruption y_{it} in municipality i, year t:

$$y_{it} = D'_{it}\beta + \delta_i + \gamma_t + \epsilon_{it} \tag{3}$$

  - D_{it}, treatments variables for years after audit
  - δ_i, municipality FE
  - γ_t, year FE

Analysis 2: Effects of auditing on subsequent corruption

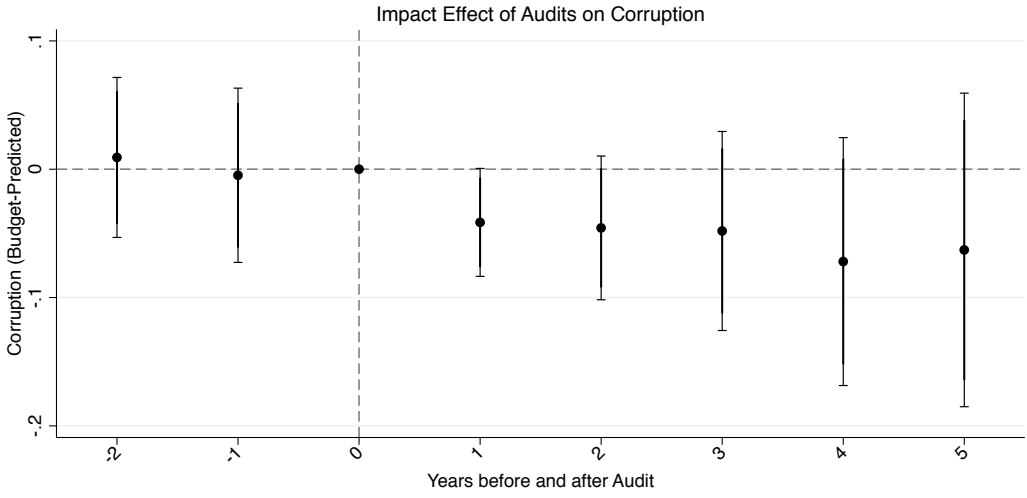
- ML-predicted corruption y_{it} in municipality i, year t:

$$y_{it} = D'_{it}\beta + \delta_i + \gamma_t + \epsilon_{it} \tag{3}$$

  - D_{it}, treatments variables for years after audit
  - δ_i, municipality FE
  - γ_t, year FE
- Empirical approach is differences-in-differences
  - What is the identification assumption for β to be consistently estimated?
  - Why is it satisfied in this case?

Event Study: Effect of Audits on Fiscal Corruption

Event Study: Effect of Audits on Fiscal Corruption

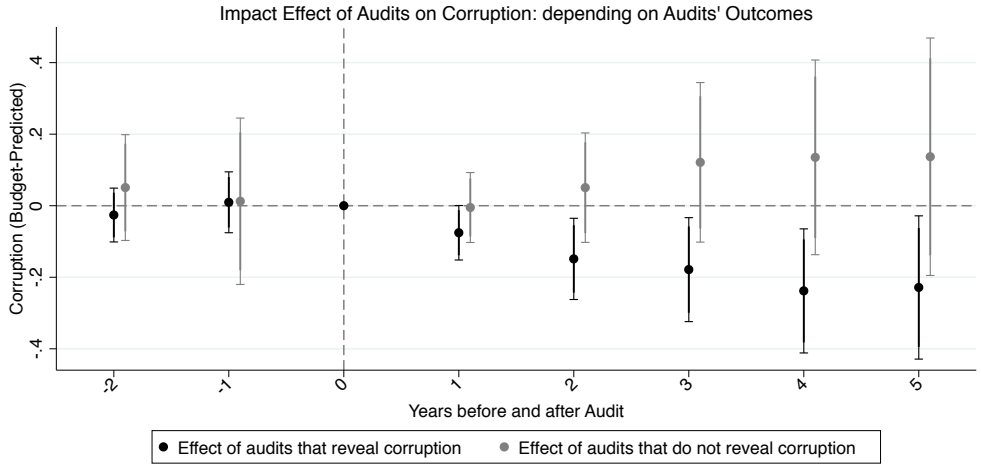


Error spikes give 95% (horizontal bars) and 90% (bold lines) confidence intervals, with standard error clustered by state.

$\Rightarrow$ The audit has a **disciplining effect**, inducing a reduction in corruption.

Event Study: By Audit Outcome

Event Study: By Audit Outcome

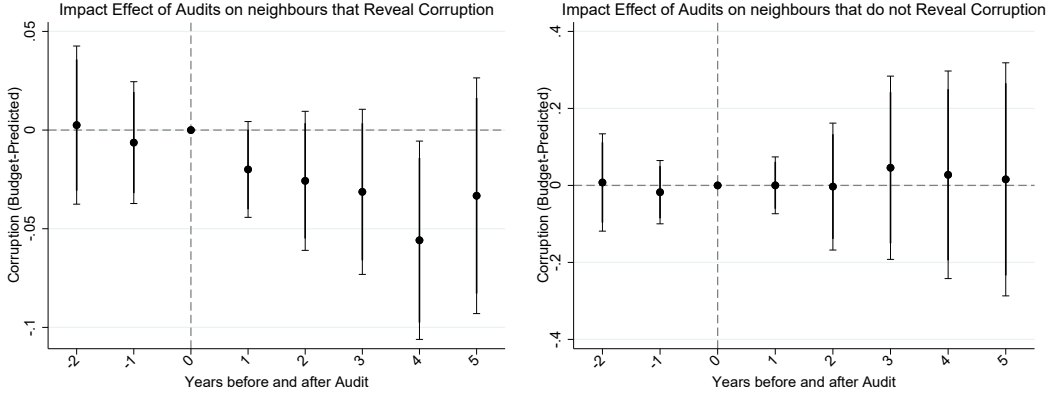


Error spikes give 95% (horizontal bars) and 90% (bold lines) confidence intervals, with standard error clustered by state.

$\Rightarrow$ When detected, fiscal corruption decreases by ~24 percentage points from a mean of 47% (approx 50 percent decrease).

Spillover Effects on Neighbors: Event Study Estimates

Spillover Effects on Neighbors: Event Study Estimates

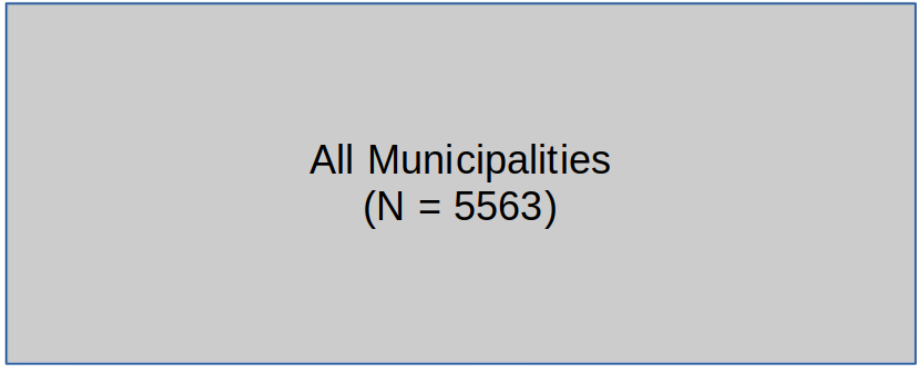


Error spikes give 95% (horizontal bars) and 90% (bold lines) confidence intervals, with standard error clustered by state.

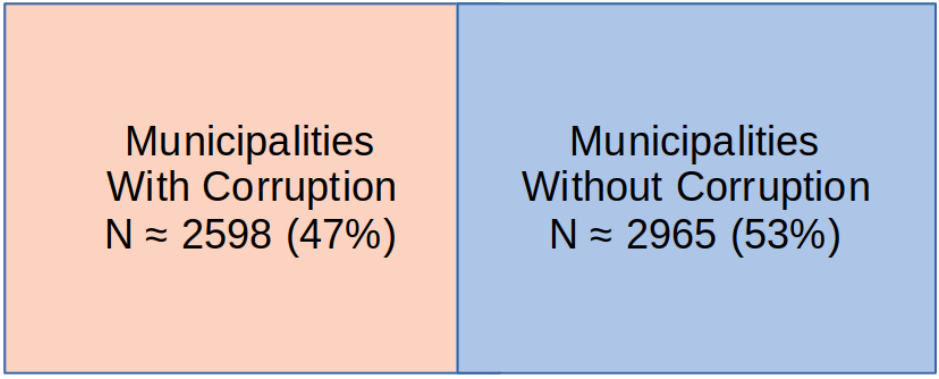
$\Rightarrow$ Effect on neighbors can be interpreted as a **behavioural response**, as audit probability is unchanged.

Outline

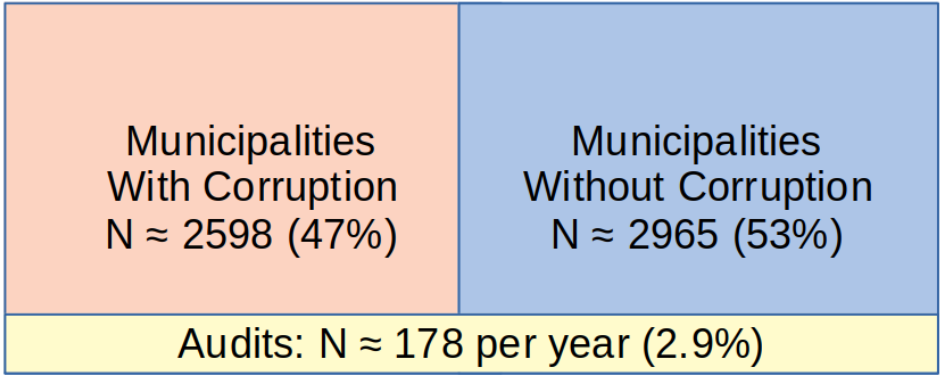
How effective are random audits?



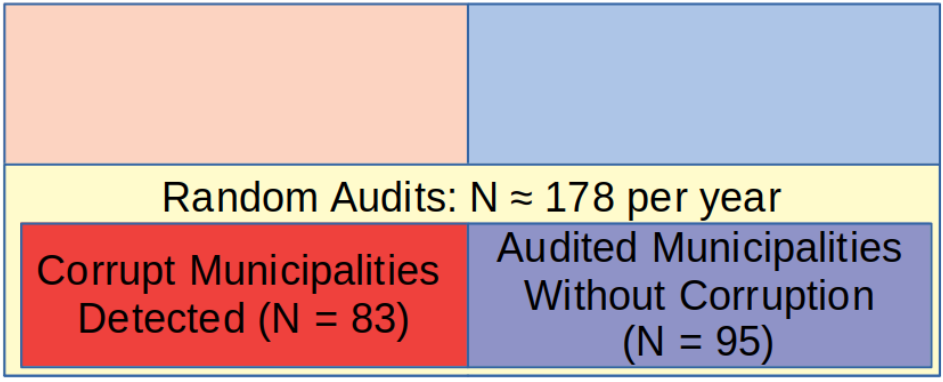
How effective are random audits?



How effective are random audits?



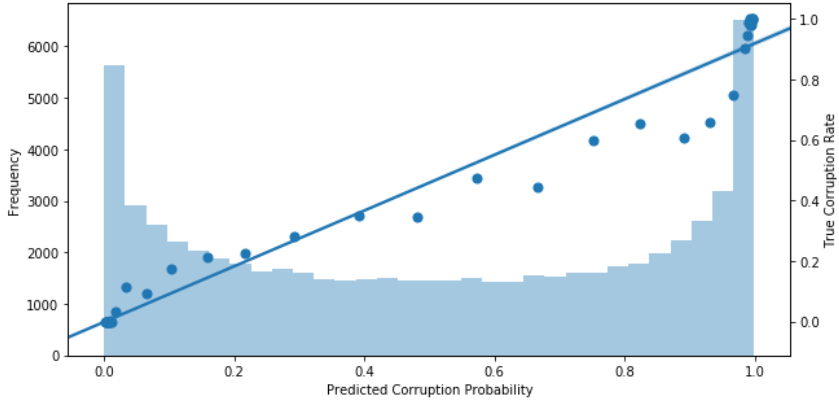
How effective are random audits?



- Under random audits, and assuming perfect detection conditional on audit, detection rate (per corrupt municipality) is equal to the audit rate (2.9%).

Targeting Audits by Corruption Risk

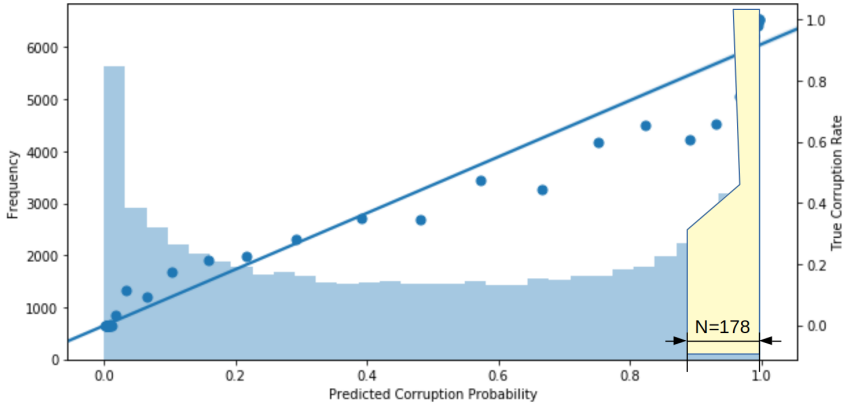
Targeting Audits by Corruption Risk



Rank municipalities by corruption risk:

- Apply model to budget data for each municipality to produce $\hat{y}_{it}$
- for each year t, get an ordinal ranking of the municipalities by predicted probability of corruption.

Targeting Audits by Corruption Risk

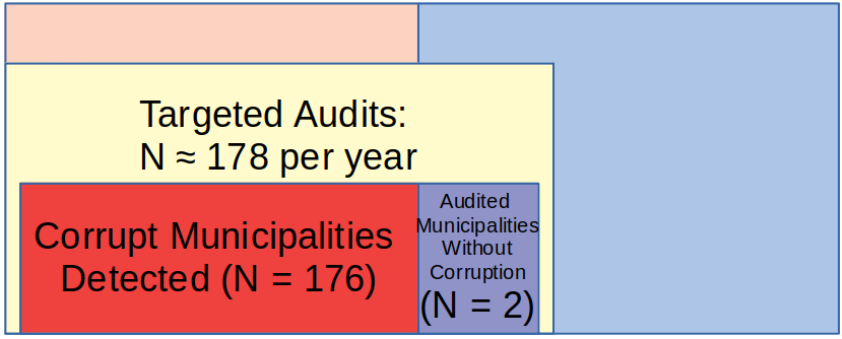


Proposed policy: Replace random audits with audits targeted by predicted corruption risk.

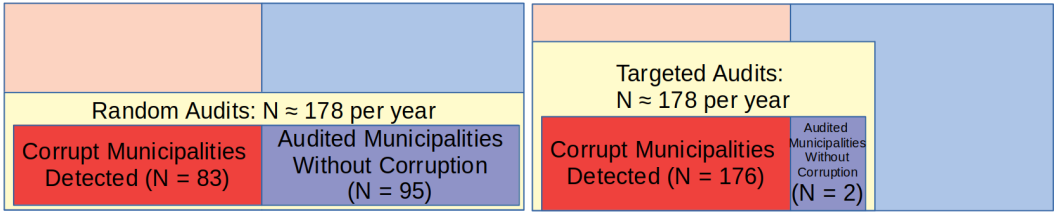
- Rather than sampling 178 municipalities uniformly from distribution, audit 178 with highest $\hat{y}_{it}$.

Targeting Audits by Corruption Risk

- ML-Targeted Auditing results in ~98% corruption detection rate.

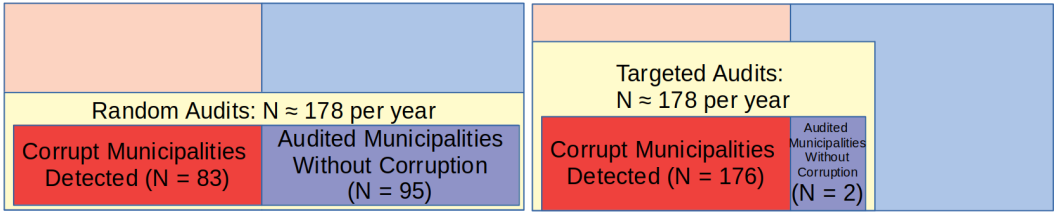


Comparing the Policies

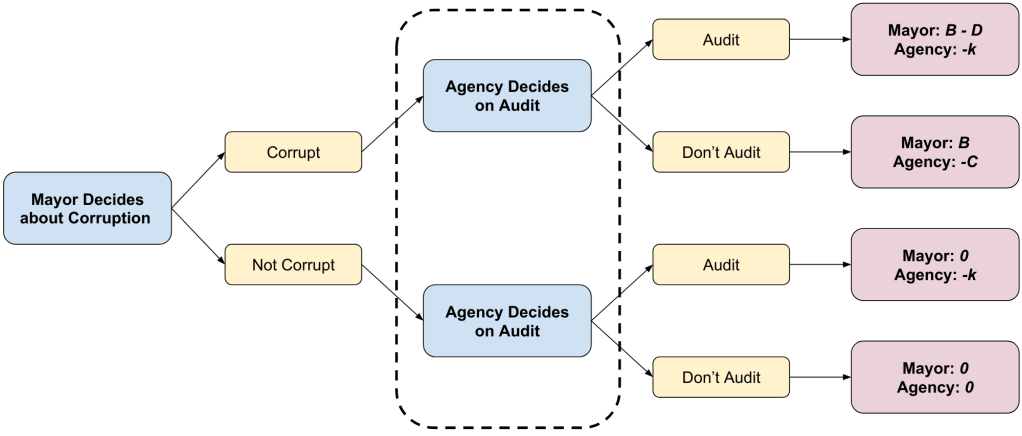


- Holding number of audits constant, targeting increases detections by 120%.
- Detection probability per corrupt municipality more than doubles – from 2.9% to 6.7%.

Comparing the Policies

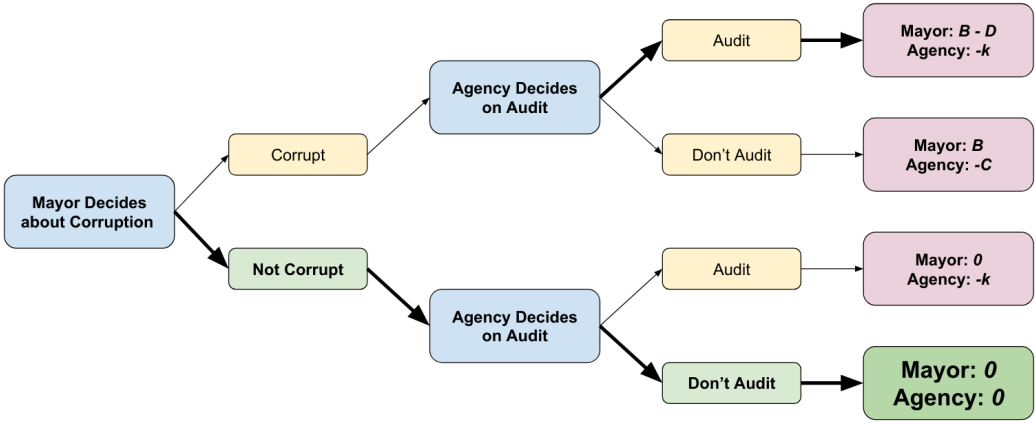


- Holding number of audits constant, targeting increases detections by 120%.
- Detection probability per corrupt municipality more than doubles – from 2.9% to 6.7%.
- To achieve same number of detections as status quo (83 municipalities), only 84 targeted audits are needed.
  - Decrease of 94 audits per year (53%), a major reduction in audit resources.
- ***Why don't we need to use the contraction method a la Kleinberg et al 2018? ("raise hand" via zoom)***



- in status quo, agency decisions are in same information set and equilibrium corruption rate is $p^* > 0$

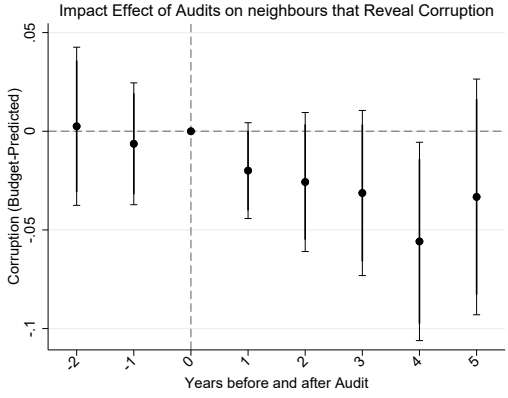
- as detection rate gets close to one, game converges to extensive form:



- by backward induction, best response is no corruption.

Behavioral AI Policy: Exploiting Spillovers

Behavioral AI Policy: Exploiting Spillovers



- According to spillover analysis, audits cut corruption by neighboring municipalities by about 10 percent (from .47 to .43) .

- Could be used to further improve policy effectiveness of targeted audits.
  - Adjust the risk ranking to target municipalities with high spillover potential.
  - For example, the policy could target the centroids of clusters of corrupt municipalities.

Recap: Brazil Corruption Study

Mechanism Design Issues

- With repeated audits, there could be behavioral responses by local officials.
  - could produce significant errors favoring savvy mayors.
  - Would still deter corrupt fiscal actions that are not easily substitutable.

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is "the industry approach", e.g., for how google/facebook detect violations.

Recap: Brazil Corruption Study

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is "the industry approach", e.g., for how google/facebook detect violations.
- mayors might learn how algorithm works over time.
- weights could be updated in response to behavioral responses

Recap: Brazil Corruption Study

Mixing random and targeted audits

- Random audits could be maintained (along with targeted audits).
  - Preserves some deterrence incentive for all municipalities.
  - Results of random audits could be used to update algorithm parameters.

Claudio Ferraz
@claudferraz

1/3 I just came across this very interesting work by @ellliottt @sergallet and @T_Giommoni using Machine Learning to predict corrupt practices in Brazil's municipalities. They show that a ML prediction algorithm can be more effective than a random auditing....

Sergio Galletta @sergallet · May 1
In a newly released WP, together with @ellliottt and @T_Giommoni, we show how ML techniques can be used to overcome data limitations when performing policy evaluation

papers.ssrn.com/sol3/papers.cf...
Show this thread

1:03 AM · Nov 29, 2020 · Twitter Web App

10 Likes

Claudio Ferraz @claudferraz · 9h
Replying to @claudferraz
2/3 But I think they miss an important point for the practical use of ML. The random audit was politically neutral and this is why it was credible to begin with. With a ML the estimated risk based on an algorithm can, in principle, be manipulated to target places or parties

1 5

Claudio Ferraz @claudferraz · 9h
3/3 So an important discussion is how to make these ML algorithms politically unbiased and how to gain credibility and convince government officials that using these types of algorithms for policy can generate important gains in the fight against corruption

What if the AI is biased toward one of the political parties?