Are Inspections Going to Waste?

Improving EPA Regulatory Compliance with Machine Learning

Michael Greenstone and Katherine Meckel

University of Chicago and University of California — San Diego

March 30, 2019

Overview

- EPA can inspect 2% of facilities for hazardous waste violations annually
  - inspector judgement, program initiatives
  - "hit rate" of uncovering severe violations ~30%
- What if we target with a predictive model?
  - 17 years of EPA data, ~10,000 variables → predict 47% improvement
  - Test of machine learning vs. decision-based targeting of gov. resources
- Field Test, FYs 2017 and 2019 → EPA, model pick 1/2 inspections
  - Prelim. results suggest 33% improvement
  - Next: scale up / across gov agencies

Overview

- EPA can inspect 2% of facilities for hazardous waste violations annually
  - inspector judgement, program initiatives
  - "hit rate" of uncovering severe violations ~30%
- What if we target with a predictive model?
  - 17 years of EPA data, ~10,000 variables → predict 47% improvement
  - Test of machine learning vs. decision-based targeting of gov. resources
- Field Test, FYs 2017 and 2019 → EPA, model pick 1/2 inspections
  - Prelim. results suggest 33% improvement
  - Next: scale up / across gov agencies

Overview

- EPA can inspect 2% of facilities for hazardous waste violations annually
  - inspector judgement, program initiatives
  - "hit rate" of uncovering severe violations ~30%
- What if we target with a predictive model?
  - 17 years of EPA data, ~10,000 variables → predict 47% improvement
  - Test of machine learning vs. decision-based targeting of gov. resources
- Field Test, FYs 2017 and 2019 → EPA, model pick 1/2 inspections
  - Prelim. results suggest 33% improvement
  - Next: scale up / across gov agencies

Background: RCRA

- Resource and Recovery Act, 1976 (RCRA)
  - regulates hazardous waste, all stages ("cradle to grave")
  - "harmful to human health or the environment"
  - mercury, petroleum, medical waste **vs.** municipal garbage, sludge
  - passed in response to, e.g., Love Canal disaster

- EPA inspects Large Quantity Generators (LQG)
  - 1,000 kgs+ of haz. waste or 1 kg+ of acutely haz. waste per month
  - industrial manufacturers, utility companies, large construction sites

Background: RCRA

- Resource and Recovery Act, 1976 (RCRA)
  - regulates hazardous waste, all stages ("cradle to grave")
  - "harmful to human health or the environment"
  - mercury, petroleum, medical waste **vs.** municipal garbage, sludge
  - passed in response to, e.g., Love Canal disaster

- EPA inspects Large Quantity Generators (LQG)
  - 1,000 kgs+ of haz. waste or 1 kg+ of acutely haz. waste per month
  - industrial manufacturers, utility companies, large construction sites

Background: Compliance Inspections

- EPA and Regions plan inspections for LQGs for FY
  - priorities, initiatives (certain areas, industries), inspector judgement
  - unannounced, walk-through, examine records
  - may occur over several days, follow-up
- Violations and Penalties:
  - corrective action, penalty, permit denial, lawsuit, criminal charges
  - (1) Mosaic Fertilizer, improper mixing of wastewater w/ corrosive substances
    - ⋆ $\rightarrow$ \$800 million in investments and penalties
  - (1) GlaxoSmithKline, improper storage of hazardous waste
  - $\rightarrow$ corrective action, penalty of \$317,550

Background: Compliance Inspections

- EPA and Regions plan inspections for LQGs for FY
  - priorities, initiatives (certain areas, industries), inspector judgement
  - unannounced, walk-through, examine records
  - may occur over several days, follow-up
- Violations and Penalties:
  - corrective action, penalty, permit denial, lawsuit, criminal charges
  - (1) Mosaic Fertilizer, improper mixing of wastewater w/ corrosive substances
    - $\rightarrow$ $800 million in investments and penalties
  - (1) GlaxoSmithKline, improper storage of hazardous waste
  - $\rightarrow$ corrective action, penalty of $317,550

Data: Outcomes

- RCRA Biennial Report, FY 2005-2017
  - All LQGs, 20,822 per year
  - Eligible for inspection: 10,407 per year
    - Manufacturing (38.0%, Chemical and Fabricated Metal), Utilities (12.4%), Transportation and Warehousing (8.9%)
    - 2.3% inspected — this is our analysis sample
    - $\rightarrow$ 35.2% find severe violation, 5-10% undetermined
    - Severe: storage w/out permit, illegal treatment or disposal, improper determination

Data: Predictive Vars

- Variables from BR and 6 other EPA programs, lagged
- LQGs regulated under Clean Air Act, TRI, etc. (more detail below)
  - historic yearly emissions, violations
- Merge together using common EPA facility ID ("registry ID")
- recode to facility-year (generally, reshape wide)
- Method for generating lags
  - Lags: **t-1**, **t-5 to t-1**, and **2000 to t-1**
  - *continuous*: mean, max, min
  - *indicators*: sum, proportion, ever 1
  - *dates*: days since most and least recent, summary
  - $\rightarrow$ 7,752 predictor variables

Data: Predictive Vars

- Variables from BR and 6 other EPA programs, lagged
- LQGs regulated under Clean Air Act, TRI, etc. (more detail below)
  - historic yearly emissions, violations
- Merge together using common EPA facility ID ("registry ID")
- recode to facility-year (generally, reshape wide)
- Method for generating lags
  - Lags: **t-1**, **t-5 to t-1**, and **2000 to t-1**
  - *continuous*: mean, max, min
  - *indicators*: sum, proportion, ever 1
  - *dates*: days since most and least recent, summary
  - $\rightarrow$ 7,752 predictor variables

RHS Variables

(1) Dataset	(2) # Init. Vars	(3) Add Lags	(4) Drop Highly Corr.	(5) Top Predictive Var, FY 2015
BR-Waste Info	57	1,681	223	Ever ship waste to Northeast, last 5 yrs
RCRA-Facility Info	374	2,489	135	Facility Latitude
RCRA-Compliance History	68	813	269	Unresolved enforcement actions
ICIS-CAA Emissions	94	340	207	"Minor Emissions" Pollutants
ICIS-Federal Enforcment	142	675	221	Date of last enforcement action, status “other”
ICIS-CWA	177	327	113	“Reconnaissance without Sampling” inspections, since 2000
TRI-Toxic Chemicals	103	1,836	368	Inspection Year * Ever Reported to TRI
Census	38	38	13	Number of Facilities in Zip Code
Other National Facility Files	24	19	5	EPA master files associate the facility with an unspecified universe
Total	1,015	7,752	1,501	

Model: Random Forest

$$viol_{it} = f(x_i, z_{t-1}, z_{(t-5 \text{ to } t-1)}, z_{(2000 \text{ to } t-1)}, u_t)$$

- $viol_{it}$ = facility i commits severe viol in year t
- time-invariant: x_i; time-varying: $z_{t-1}, z_{t-5 \text{ to } t-1}, z_{2000 \text{ to } t-1}$; time FE: u_t

- Random Forest averages over many decision trees (Breiman, 2001)
  - Classification and Regression Tree (**CART**), 0-1 outcome
  - at node, find predictor and "split value" to minimize error
    - minimize $\sum_{c=1}^{2} -\bar{y}_c(1-\bar{y}_c)$, c denotes child set, $\bar{y}_c = \text{mean}(y)$ for $y \in c$
  - **highly flexible**, considers all non-linearities and interactions, but suffers from **overfit**
    - $\rightarrow$ draws random subset of obs. (2/3) and vars (square root)
    - $\rightarrow$ average prediction over trees, reduce fit to idiosync.
    - does well relative to other ML algorithms on flexibility and reducing overfit

Model: Random Forest

$$viol_{it} = f(x_i, z_{t-1}, z_{(t-5 \text{ to } t-1)}, z_{(2000 \text{ to } t-1)}, u_t)$$

- $viol_{it}$ = facility i commits severe viol in year t
- time-invariant: x_i; time-varying: $z_{t-1}, z_{t-5 \text{ to } t-1}, z_{2000 \text{ to } t-1}$; time FE: u_t

- Random Forest averages over many decision trees (Breiman, 2001)
  - Classification and Regression Tree (**CART**), 0-1 outcome
  - at node, find predictor and "split value" to minimize error
    - ★ minimize $\sum_{c=1}^{2} -\bar{y}_c(1-\bar{y}_c)$, c denotes child set, $\bar{y}_c = \text{mean}(y)$ for $y \in c$
  - **highly flexible**, considers all non-linearities and interactions, but suffers from **overfit**
    - ★ $\rightarrow$ draws random subset of obs. (2/3) and vars (square root)
    - ★ $\rightarrow$ average prediction over trees, reduce fit to idiosync.
    - ★ does well relative to other ML algorithms on flexibility and reducing overfit

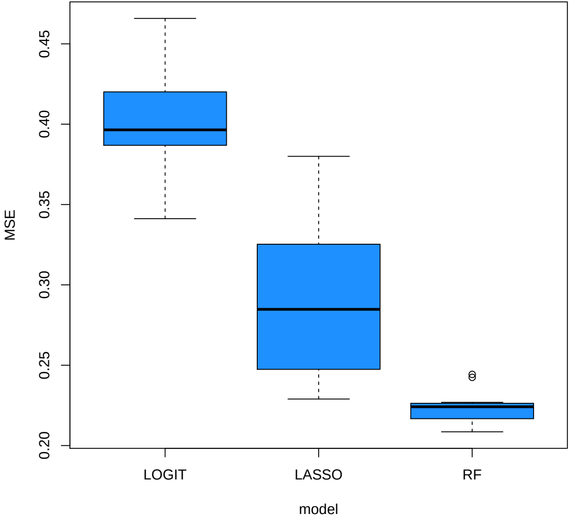
Model: Random Forest

$$viol_{it} = f(x_i, z_{t-1}, z_{(t-5 \text{ to } t-1)}, z_{(2000 \text{ to } t-1)}, u_t)$$

- $viol_{it}$ = facility i commits severe viol in year t
- time-invariant: x_i; time-varying: $z_{t-1}, z_{t-5 \text{ to } t-1}, z_{2000 \text{ to } t-1}$; time FE: u_t

- Random Forest averages over many decision trees (Breiman, 2001)
  - Classification and Regression Tree (**CART**), 0-1 outcome
  - at node, find predictor and "split value" to minimize error
    - ★ minimize $\sum_{c=1}^{2} -\bar{y}_c(1-\bar{y}_c)$, c denotes child set, $\bar{y}_c = \text{mean}(y)$ for $y \in c$
  - **highly flexible**, considers all non-linearities and interactions, but suffers from **overfit**
    - ★ $\rightarrow$ draws random subset of obs. (2/3) and vars (square root)
    - ★ $\rightarrow$ average prediction over trees, reduce fit to idiosync.
    - ★ does well relative to other ML algorithms on flexibility and reducing overfit

MSE, Different Predictive Models, FY 2017



Performance Metric for Model Parameter Choices

- Simulate field test
  - Estimate RF for $2005 - t$, generate risk scores eligible LQGs in $t + 1$, calculate hit rate on top 2%
  - Repeat for $t = 2010, 2011..., 2014$, average hit rates across 2011-2015
  - caveat: hit rate missing if facility not inspected

Model Choices

- RF prediction error increases in correlation between trees
  - *nearZeroVar* from **caret** pkg
    - removes extremely low var features
  - *findCorrelation* from **caret** pkg
    - removes one of high corr. variables
  - Positive importance
    - drop variables with importance less than 0.1 in initial model run
  - $\rightarrow$ 1,501 out of about 10,000 vars

Model Choices

- m: total variables sampled per node
  - no marginal improvements away from $39 = \sqrt{1501}$
- Propensity score weighting (draws proportional to bins)
  - adjust for selection into inspection
- Class imbalance adjustments (interested in minority class)
- Model's inspections uncover **47.4%** more violations than EPA's
  - 38% to 56%

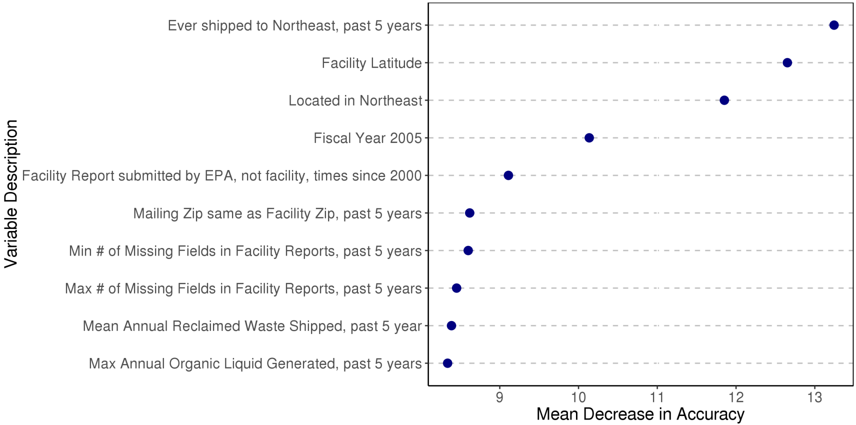
Model Choices

- m: total variables sampled per node
  - no marginal improvements away from $39 = \sqrt{1501}$
- Propensity score weighting (draws proportional to bins)
  - adjust for selection into inspection
- Class imbalance adjustments (interested in minority class)
- Model's inspections uncover **47.4**% more violations than EPA's
  - 38% to 56%

Model Precision by Included Features, FY 2017

Drop	# of features	OOB Error	N Top 5%	N Top 10%	Precis. 5%	Precis. 10%
Pos. Imp	1,421	36.1	117	200	57.3	55.5
Zero Var	4,163	36.1	125	203	56.0	51.2
Highly Corr.	1,624	35.5	127	203	52.8	50.7
Manual	1,554	35.9	125	202	55.2	50.5
Insp. Wts	1,554	36.1	115	206	53.0	51.9

Top 10 Variables, by Importance, FY 2017



FY17 Field Test: Eligible Facilities and Inspection Targets

Region	Elig. Facilities		Requested		Committed	
	Original	Final	EPA	Model	EPA	Model
1	578	345	5	6	4	4
2	2356	242	36	37	36	37
3	744	628	10	11	10	5
4	675	258	19	19	19	19
5	1690	732	25	25	25	25
6	1016	915	5	5	-	-
7	288	43	16	16	6	3
8	60	22	14	15	11	11
9	1400	1399	10	10	10	10
10	210	18	8	8	-	-
Total	9017	4602	148	152	121	114

FY19 Field Test: Eligible Facilities and Inspection Targets

Region	Elig. Facilities		Requested		Committed	
	Original	Final	EPA	Model	EPA	Model
1	629	-	-	-	-	-
2	2633	201	30	30	30	30
3	621	621	6	6	7	7
4	375	92	10	10	6	6
5	1381	963	25	25	25	25
6	1200	-	-	-	-	-
7	81	53	6	6	4	4
8	85	42	5	5	8	8
9	1794	1544	10	10	14	14
10	218	-	-	-	-	-
Total	9017	3516	92	92	94	94

Field Test Results

	(1)	(2)	(3)	(4)
Model Pick	0.1194**	0.0691	0.0713	
	(0.0589)	(0.0601)	(0.0955)	
Region=1 × Model Pick				0.0000
				(0.0000)
Region=2 × Model Pick				0.0989
				(0.0678)
Region=3 × Model Pick				0.0417
				(0.3371)
Region=4 × Model Pick				0.4059**
				(0.1911)
Region=5 × Model Pick				0.2419+
				(0.1253)
Region=7 × Model Pick				-0.8333***
				(0.1595)
Region=9 × Model Pick				-0.0281
				(0.2040)
Mean, Dep. Var.	0.2950	0.2950	0.2950	0.2950
Fixed Effects	R*W	R*W, Insp.	R*W	R*W
Weights	No	No	Yes	No
Cluster	Fac*Wave	Fac*Wave	Fac*Wave	Fac*Wave
Observations	200	200	200	200