nature sustainability

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41893-018-0142-9

In the format provided by the authors and unedited.

Machine learning for environmental monitoring

M. Hino, E. Benami* and N. Brooks

Stanford University, Stanford, CA, USA. *e-mail: elinor@stanford.edu

Supplementary Information for

Machine learning for environmental monitoring

M. Hino, E. Benami, N. Brooks

Corresponding Author: E. Benami
Email: elinor@stanford.edu

This PDF file includes:

Supplementary Notes

Supplementary Figures

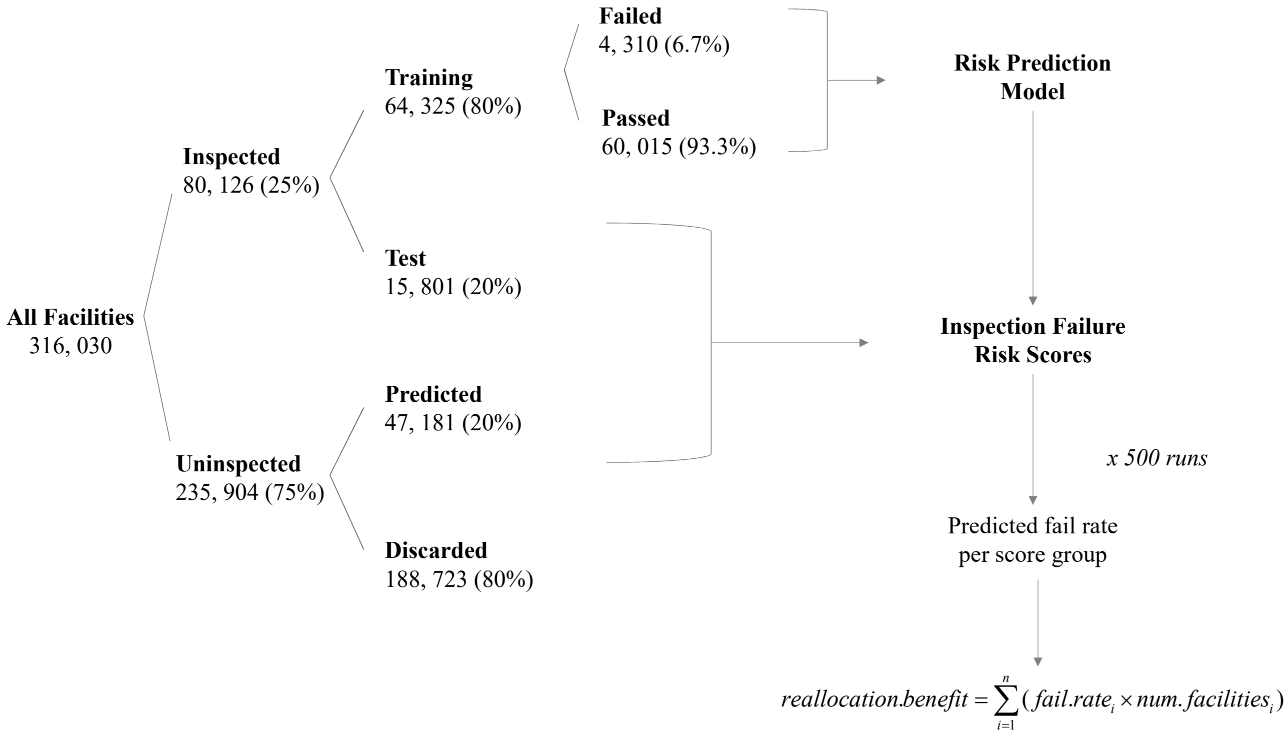
Supplementary Tables

Supplementary Notes

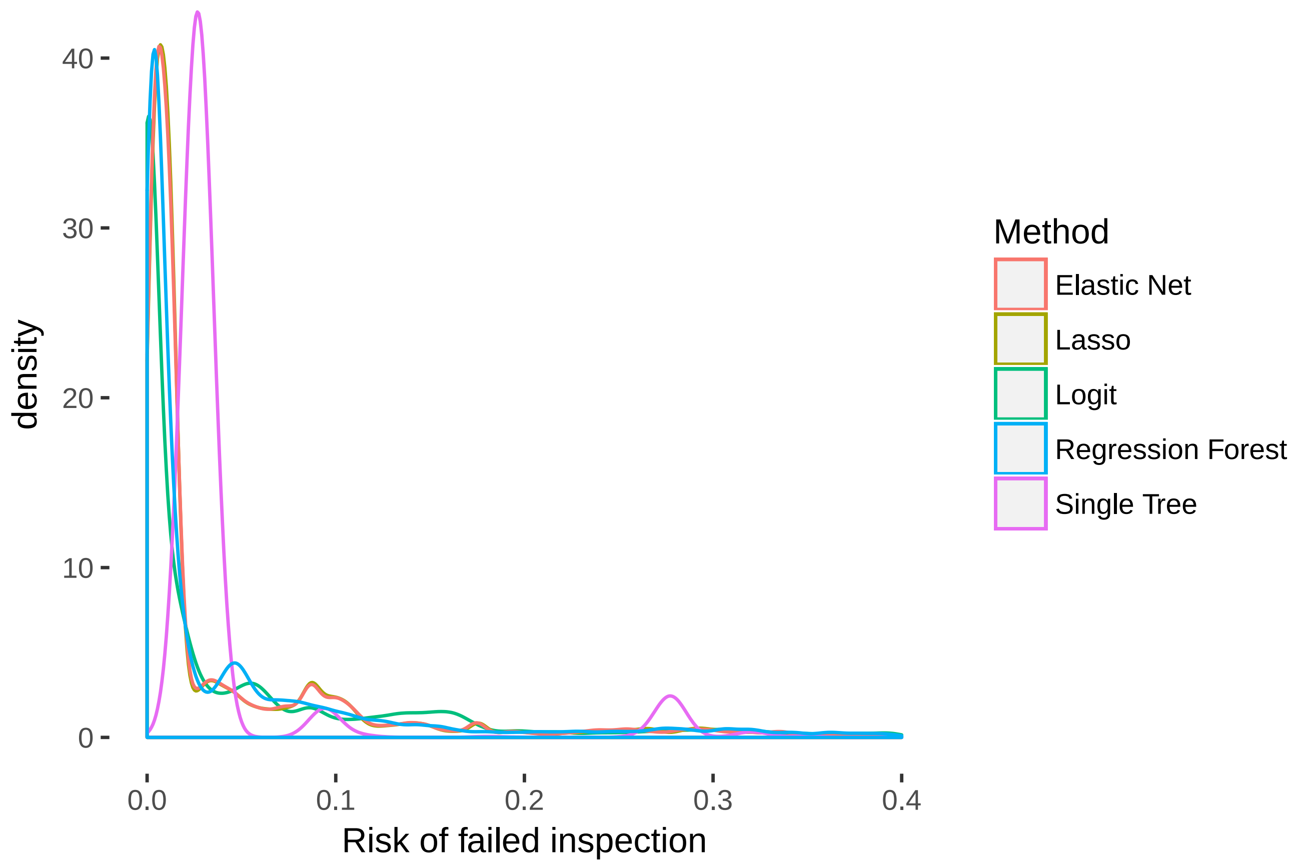
Note 1. Estimates for inspection costs

$20M stems from authors' estimates of 1 person per inspection at a $50,000/year salary (rounding up from salary estimates available from the U.S. Bureau of Labor Statistics on for Environmental Science and Protection Technicians [1]) and 26 hours of preparation, travel, and write-up time — totaling approximately $620/inspection across $\sim$ 36,500 inspections per year (distributed among 27,500 unique facilities). Other sources have estimated total inspection costs of up to $40 million based on their own calculations for Clean Air Act inspections [2], which could differ from the Clean Water Act.

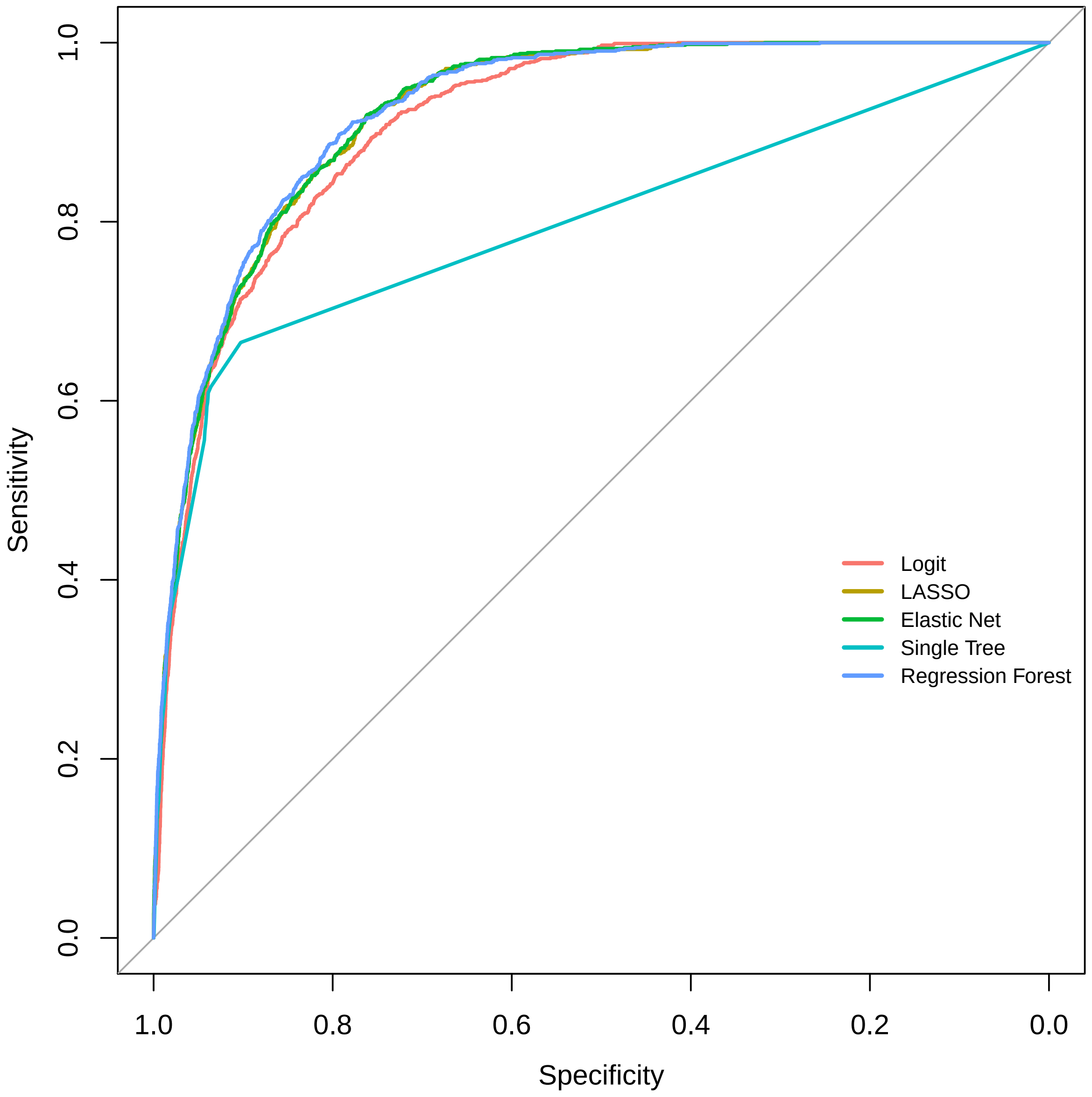
Supplementary Figures



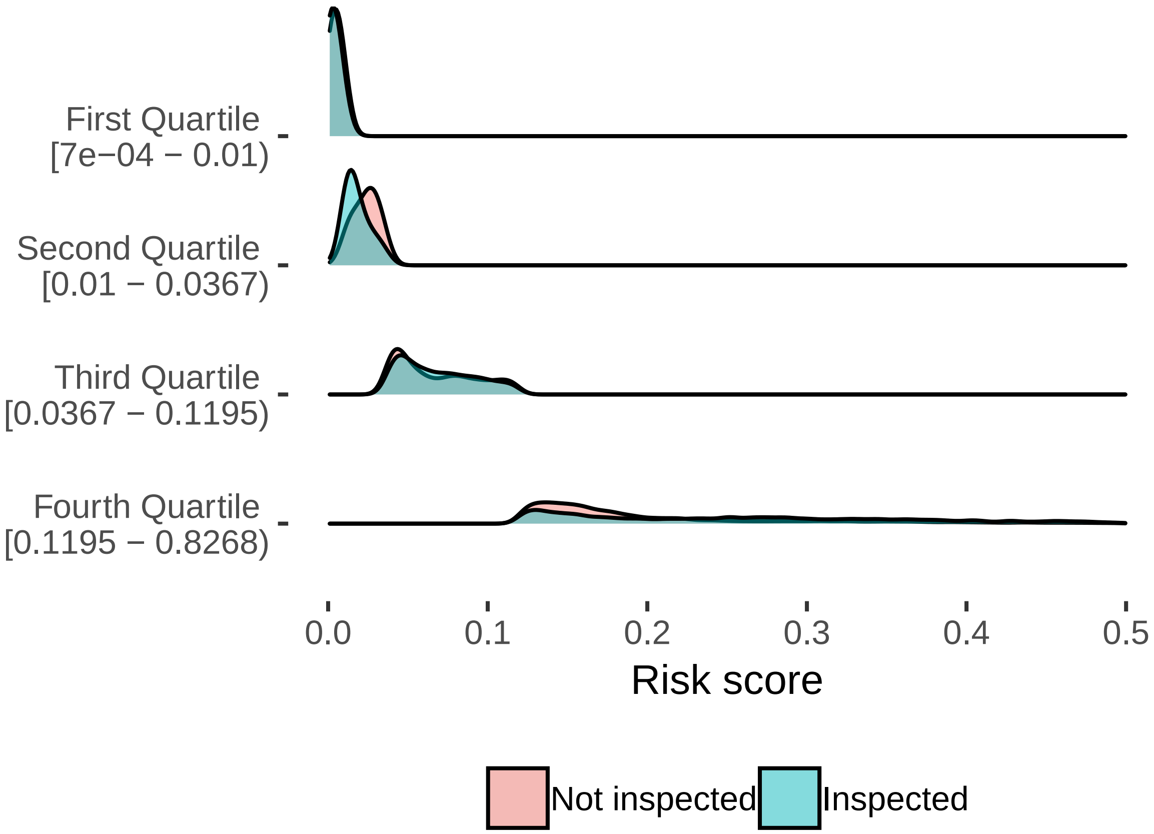
Supplementary Figure 1: Schematic for data flow & use. ECHO provides information on $\sim$ 393,220 CWA-regulated facilities. We filter this to a set of 316,030 for which we have all covariate information. We then split the data based on inspection history and divide the inspected data into an 80% training and 20% test set. The training set is discarded after developing the model, and a 20% inspected test dataset plus a 20% random sample of uninspected facilities are used for the reallocation results described in this paper.



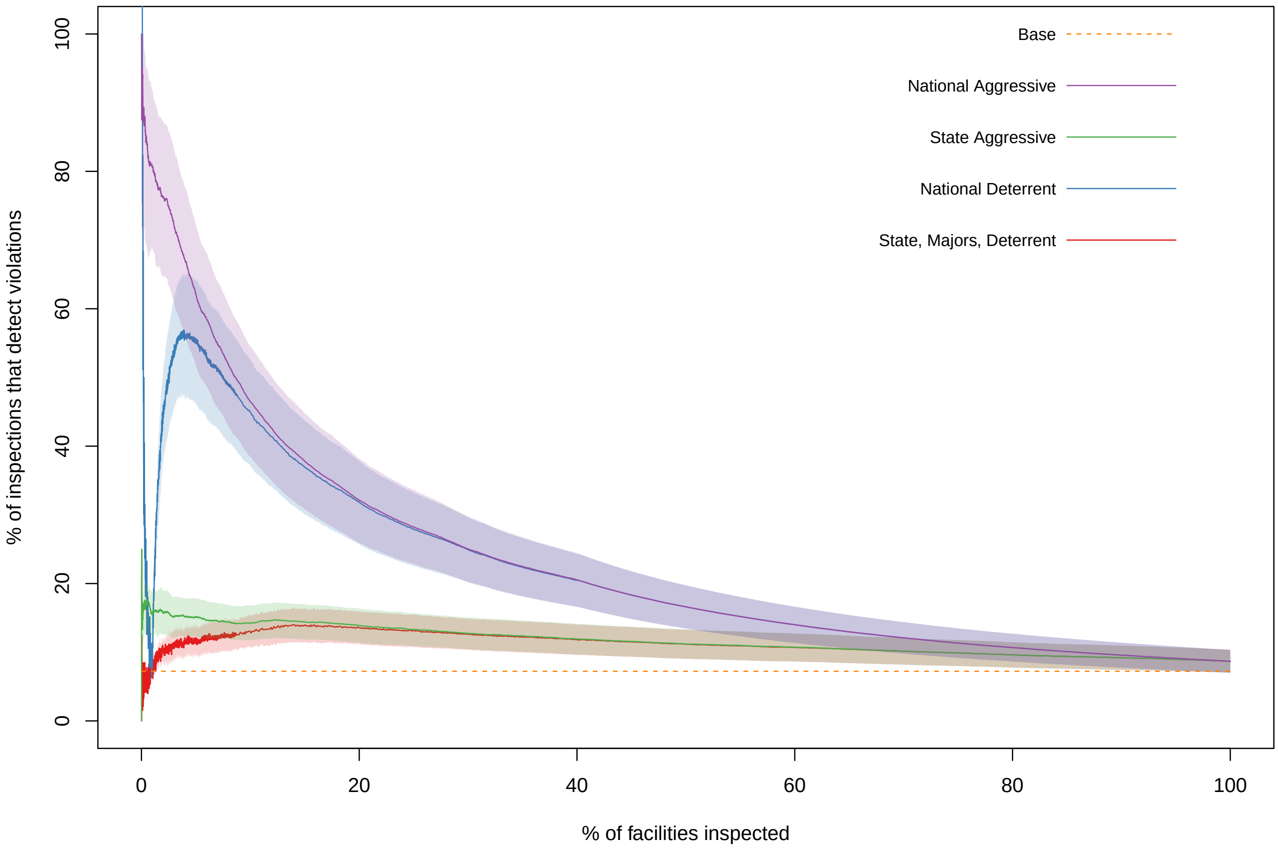
Supplementary Figure 2: Risk score distribution by method. The risk scores for nearly all methods cluster between 0 and 0.1. The density plots also show that the single tree exhibits more variation in resulting risk scores than any other model. It also exhibits poorer performance than any other model, when evaluated by the AUC metric, as shown in Supplementary Figure 3 and Supplementary Table 3.



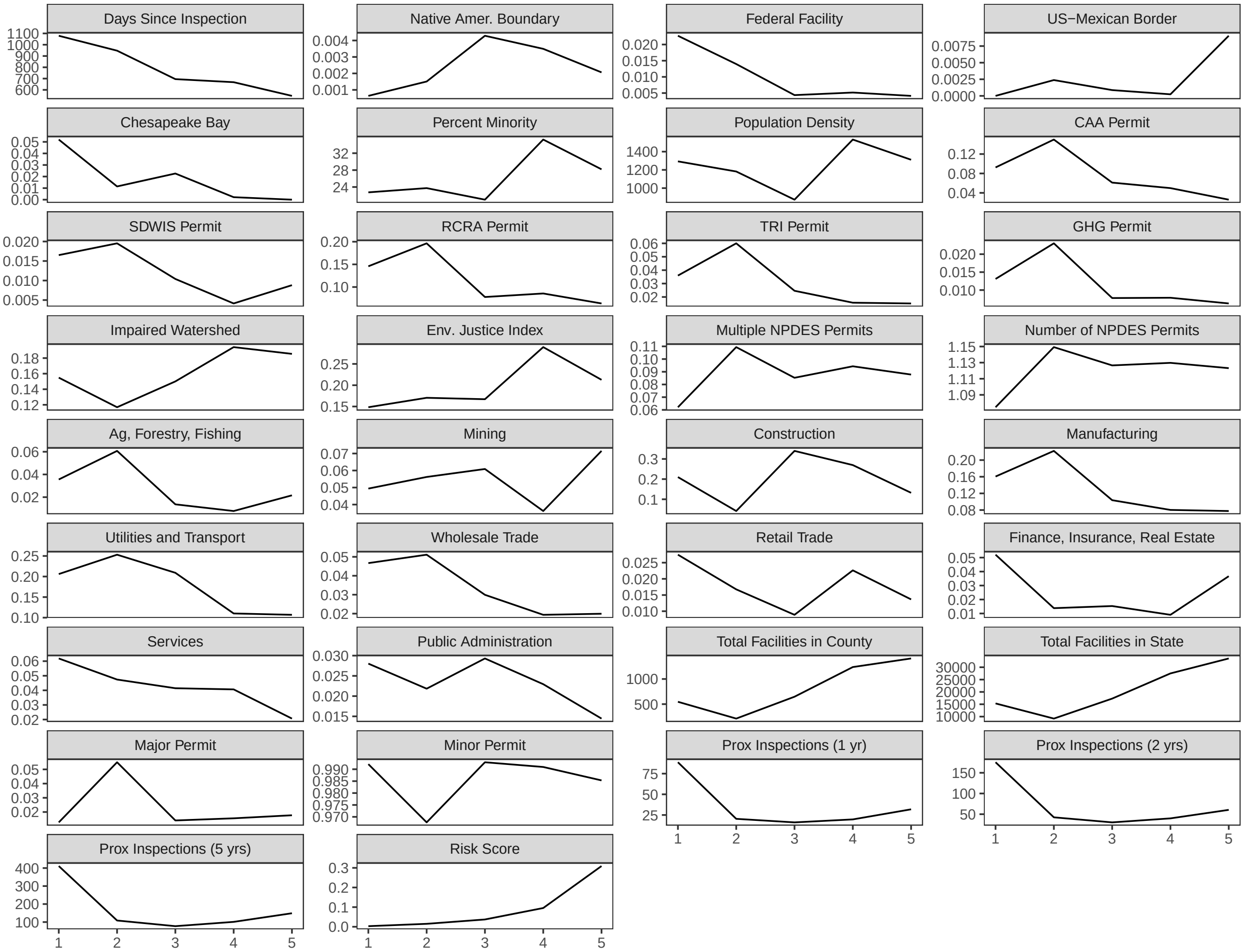
Supplementary Figure 3: Receiver Operating Characteristic (ROC) Curve. An ROC curve maps the trade-off between the false positive rate (specificity, type-II error, x-axis) against the true positive rate (sensitivity, type-I error, y-axis). Curves closer to the northwestern corner possess a higher area under the curve (AUC), indicating better performance [3]. The x-axis represents the specificity, i.e., false positive rate, and the y-axis presents the sensitivity, i.e., true positive rate. Nearly all methods appear to perform similarly, except for the single tree, which performs substantially worse than all the rest. The dotted 45 degree line represents the "no information" classifier, i.e., what we would expect if facility characteristics were not associated with the probability of failing an inspection, which has an AUC of 0.5.



Supplementary Figure 4: Distribution of risk scores by inspection status. Most facilities are clustered at very low risk scores. Examining the distribution by inspected vs. uninspected, we find substantial overlap between the two sets of facilities; that is, it is not the case that all of the inspected facilities have high risk scores and the uninspected have low risk scores. Note that 3% of facilities have risk scores above 0.5, though these have been truncated to permit visualization.



Supplementary Figure 5: Violation rates as a function of the number of inspections. The percentage of inspections that detect violations depends in part on how many inspections total are allocated: as more facilities are assigned to inspection, more lower-risk facilities are inspected, and the predicted share of inspections that detect violations declines. As shown here, as resources become scarcer, the value of targeting inspections through machine learning increases. If all facilities can be inspected, there is no targeting, and therefore no additional benefit. The dashed orange line shows the current share of inspections that detect violations, and the solid lines show the percentage of inspections that detect violations under different allocations. The semi-transparent ribbons around each line indicate the 95% confidence interval. The “deterrence” allocations (blue and red lines) move non-monotonically because at low numbers of inspections, all inspections are dedicated to the random 1% draw of facilities. As inspections increase, more inspections can be assigned based on risk rather than deterrence.



Supplementary Figure 6: Risk scores vs. facility characteristics. Examining how each of the included covariates varies as a function of risk scores can shed insight into which characteristics are influential in predicting the likelihood of failing CWA inspections. Each graph above shows mean covariate values evaluated at each quintile of risk scores. The risk scores defining the quintiles are: [0.00073,0.0073], (0.0073,0.026], (0.026,0.053], (0.053,0.142], (0.142,0.827]. For binary variables, the y-axis represents the proportion of facilities in the quintile that meet the criteria, e.g., the number that fall within the Mining industry classification or the number that have multiple NPDES permits. Note a single facility can be categorized under multiple industries. Facilities in the manufacturing, utilities and transportation, and wholesale trade industries, for example, are more concentrated in the lower-risk quintiles than the higher-risk quintiles. On the other hand, classification in the retail trade sector shows little correlation with risk score. Facilities that also have permits under other environmental regulations, such as the Clean Air Act and the Resource Conservation and Recovery Act, appear less likely to fail an inspection, with a greater share of facilities in low-risk quintiles. This figure is produced with the combined 20% inspected test dataset plus a 20% random sample of uninspected facilities (see Supplementary Figure 1).

Supplementary Tables

Covariate Name	**Type**	**Description**
State	Factor	Levels for states, District of Columbia, and territories
Chesapeake Bay	Binary	In Chesapeake Bay watershed
Tribal Boundary	Binary	In Native American tribal boundary
US-Mexico border	Binary	Within 100km of US-Mexico border
Percent minority	Continuous	% nonwhite population within a 3-mile radius
Population density	Continuous	# of people per square mile within a 3-mile radius
Major facility	Binary	Designated as a "major" **
Minor facility	Binary	Designated as a "minor" **
Multiple permits	Binary	Has multiple NPDES permits
Number of permits	Continuous	Number of NPDES permits
Air	Binary	Regulated under Clean Air Act
Safe Drinking Water	Binary	Regulated under Safe Drinking Water Act
RCRA	Binary	Regulated under Resource Conservation & Recovery Act
TRI	Binary	Registered to report to Toxics Release Inventory
Greenhouse Gas	Binary	Regulated under the GHG Reporting Program
Impaired water	Binary	Discharges into designated impaired waters
EJ Screen	Binary	In a census block group that is $\geq$ 80th percentile in an index of EPA's enviro. justice screening tool, EJSCREEN
County facilities*	Continuous	# CWA regulated facilities in same county
State facilities*	Continuous	# CWA regulated facilities in same state/territory
Party*	Factor	Governor political party: Democrat, Independent, or Republican
Agriculture	Binary	Industry code A: Ag., Forestry, & Fishing
Mining	Binary	Industry code B: Mining
Construction	Binary	Industry code C: Construction
Manufacturing	Binary	Industry code D: Manufacturing
Transport. & Utilities	Binary	Industry code E: Transportation, Communications, Electric, Gas, & Sanitary Services
Wholesale trade	Binary	Industry code F: Wholesale Trade
Retail trade	Binary	Industry code G: Retail Trade
Finance	Binary	Industry code H: Finance, Insurance, & Real Estate
Services	Binary	Industry code I: Services
Public Admin.	Binary	Industry code J: Public Administration
Time since last inspection*	Continuous	Days elapsed since prior inspection
Proximate inspections*	Continuous	# CWA regulated facilities within 50 miles inspected in the year prior to inspection
DMRs submitted*	Continuous	# DMRs submitted in year prior to inspection
DMR E90*	Continuous	# DMRs submitted that report effluent violation in year prior to inspection
DMR D80*	Continuous	# DMRs submitted late in year prior to inspection (no numeric limit)
DMR D90*	Continuous	# DMRs submitted late in year prior to inspection (with numeric limit)

Supplementary Table 1: Covariate names and descriptions. This table describes each variable used in our analysis. All binary variables are coded as one if the description is true and zero otherwise. One asterisk (*) indicates a variable constructed for analysis by the authors. Two asterisks (**) flags that a facility can be classified as both major and minor if it has multiple permits. All variables are drawn or derived from publicly available databases on EPA's website, with the exception of data on the governor's political party for the state each facility is located in, which was obtained from the National Governors Association [4]. We interact all variables for the elastic net and LASSO models.

Variable Name	Mean	SD	(Min - Max)
Total Facilities in State	20,542	17,053.9	(10 - 55,115)
Population Density	1,228	2,261.6	(0 - 57,168)
Total Facilities in County	809	1,079.0	(1 - 10,544)
Proximate Inspections (5 yrs)	168	288.4	(0 - 3,309)
Proximate Inspections (2 yrs)	69	120.5	(0 - 1,368)
Proximate Inspections (1 yr)	35	61.0	(0 - 726)
Percent Minority	26	23.4	(0 - 100)
Number of NPDES Permits	1	1.0	(1 - 390)
Days Since Inspection	797.33	1355.01	(1 - 14270)
Chesapeake Bay	0.00	0.05	(0 - 1)
Native Amer. Boundary	0.01	0.10	(0 - 1)
US-Mexican Border	0.00	0.05	(0 - 1)
Federal Facility	0.02	0.13	(0 - 1)
CAA Permit	0.08	0.26	(0 - 1)
SDWIS Permit	0.01	0.11	(0 - 1)
RCRA Permit	0.11	0.32	(0 - 1)
TRI Permit	0.03	0.17	(0 - 1)
GHG Monitoring	0.01	0.11	(0 - 1)
Impaired Watershed	0.16	0.37	(0 - 1)
Env. Justice Index	0.20	0.40	(0 - 1)
Multiple NPDES Permits	0.09	0.28	(0 - 1)
Ag, Forestry, Fishing	0.03	0.16	(0 - 1)
Mining	0.05	0.23	(0 - 1)
Construction	0.20	0.40	(0 - 1)
Manufacturing	0.13	0.33	(0 - 1)
Utilities and Transport	0.18	0.38	(0 - 1)
Wholesale Trade	0.03	0.18	(0 - 1)
Retail Trade	0.02	0.13	(0 - 1)
Finance, Insurance, Real Estate	0.03	0.16	(0 - 1)
Services	0.04	0.21	(0 - 1)
Public Administration	0.02	0.15	(0 - 1)
Major Permit	0.02	0.15	(0 - 1)
Minor Permit	0.99	0.12	(0 - 1)

Supplementary Table 2: Descriptive statistics for covariates of complete dataset (n = 316,030). This table provides descriptive statistics for all of the variables used in our analysis.

	Sensitivity (TPR)	Specificity (TNR)	Accuracy	Precision (PPV)	Negative Predictive Value (NPV)	AUC
Logit	13.6%	99.2%	93.6%	53.3%	94.2%	0.907
LASSO	31.2%	98.4%	94%	58%	95.3%	0.918
Elastic Net	31.2%	98.4%	94%	58.1%	95.3%	0.918
Single Tree	25.4%	98.9%	94.1%	61.6%	95%	0.782
Regression Forest	25.7%	98.9%	94.1%	62.9%	95%	0.922

Supplementary Table 3: Comparison of prediction metrics by model. This table reports a standard set of prediction metrics, including the sensitivity (true positive rate), specificity (true negative rate), accuracy (rate of correct predictions out of all predictions), precision (or positive predictive value, which reflects the rate of true positives out of all positive predictions), negative predictive value (true negative predictions out of all negative predictions), and AUC. The AUC is the area under the Receiver Operating Characteristic (ROC) curve, which is further explored in Supplementary Figure 3. By comparing the metrics in this table across models, we can evaluate the performance of each model in order to select the one that yields the best predictions. We also computed the same metrics while varying the threshold from 0.5 to 0.2 (results not shown but available from authors on request). As we lower the threshold for converting the continuous risk score to a binary pass/failure, we increase the sensitivity (true positives): Lowering the threshold increases the number of true positives for all models, but for all thresholds we tested, the regression forest remained the best performing model. Lowering the threshold had little influence on the accuracy, which remained above 90% for all methods and is driven by the high predictive ability of our models to correctly identify true negatives.

	MSE w/o DMR data	MSE w/DMR data	AUC w/o DMR data	AUC w/DMR data
Logit	0.0527	0.0514	0.9270	0.9300
Lasso	0.0521	0.0502	0.9260	0.9300
Elastic Net	0.0520	0.0502	0.9260	0.9290
Single Tree	0.0547	0.0537	0.8660	0.9020
Regression Forest	0.0520	0.0503	0.9310	0.9350

Supplementary Table 4: MSE and AUC comparison with and without self-reported data. This table compares the performance of the five different machine learning models on the subset of facilities that submit Discharge Monitoring Reports (DMRs), evaluated with and without the additional information provided by the DMRs. The regression forest performs the best with and without the DMR data using the AUC criterion.

	Share of inspections that detect failures	95% confidence interval	Improvement over base case (% change)
Aggressive, national	0.50	(0.42, 0.58)	639
Aggressive, state	0.15	(0.13, 0.17)	119
Deterrence, national	0.48	(0.4, 0.56)	608
Majors, deterrence, state	0.13	(0.11, 0.15)	92

Supplementary Table 5: Expected share of inspections that detect failures and the 95% confidence interval per reallocation proposal. All four reallocations substantially increase the share of inspections that detect failures. Rates are calculated by reallocating 20% of average annual inspections (5,480) among the 20% test case dataset (n = 15,801) and a 20% random sample of the uninspected facilities (n = 47,181) per the rules of each reallocation proposal.

	Inspected and uninspected	Only Inspected
Aggressive, national	0.5 (0.42 - 0.58)	0.47
Aggressive, state	0.15 (0.13 - 0.17)	0.14
Deterrence, national	0.48 (0.4 - 0.56)	0.44
Majors, deterrence, state	0.13 (0.11 - 0.15)	0.11

Supplementary Table 6: Results of reallocations over only inspected facilities. Using a proportionate number of inspections on only the inspected facilities, the improvements in the number of violations detected are comparable to the improvements when reallocating over inspected and uninspected facilities. Column 1 is identical to the results shown in Supplementary Table 5 and are reported again here only for direct comparison to the sample of only-inspected facilities (Column 2).

References

[1] U.S. Bureau of Labor Statistics. Environmental Science and Protection Technicians: Occupational Outlook Handbook. https://www.bls.gov/ooh/life-physical-and-social-science/environmental-science-and-protection-technicians.htm. (Accessed on 07/14/2017).

[2] Rema Nadeem Hanna, Paulina Oliva, et al. "The Impact of Inspections on Plant-Level Air Emis-sions". In: The BE Journal of Economic Analysis & Policy 10.1 (2010), pp. 1–29.

[3] Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani. An Introduction to Statistical Learning. (2013). An Introduction to Statistical Learning. New York: Springer.

[4] National Governors Association. National Governors Association. https://www.nga.org/. (Ac-cessed on 06/13/2017).