nature sustainability

ARTICLES

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

Machine learning for environmental monitoring

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

Public agencies aiming to enforce environmental regulation have limited resources to achieve their objectives. We demonstrate how machine-learning methods can inform the efficient use of these limited resources while accounting for real-world concerns, such as gaming the system and institutional constraints. Here, we predict the likelihood of a facility failing a water-pollution inspection and propose alternative inspection allocations that would target high-risk facilities. Implementing such a data-driven inspection allocation could detect over seven times the expected number of violations than current practices. When we impose constraints, such as maintaining a minimum probability of inspection for all facilities and accounting for state-level differences in inspection budgets, our reallocation regimes double the number of violations detected through inspections. Leveraging increasing amounts of electronic data can help public agencies to enhance their regulatory effectiveness and remedy environmental harms. Although employing algorithm-based resource allocation rules requires care to avoid manipulation and unintentional error propagation, the principled use of predictive analytics can extend the beneficial reach of limited resources.

Machine-learning techniques are a potentially valuable tool for public agencies seeking to employ 'big data' to inform decision-making[1–3]. By generating data-based predictions, these techniques enable scarce resources to be allocated to maximize potential benefits. Machine-learning approaches have been applied in diverse policy contexts, such as predicting crime risk among pre-trial criminal defendants[4] and public health code violations among restaurants[5], and identifying who will benefit most from at-risk youth interventions[6].

However, applying machine learning to public policy problems can present new challenges. First, the use of machine-learning approaches may induce strategic behaviour. For example, agents may manipulate their reported data to influence the likelihood of receiving benefits or avoiding penalties[2]. In addition, even if agents do not intentionally misrepresent their own data, they may alter their behaviour if they know their likelihood of being 'selected' by the algorithm. For example, if an actor learns they face low risk of some potentially punitive measure, they might relax their standards. Data-based predictions may also codify historical human biases or institutionalize systematic gaps in data coverage, and therefore lead to unintended consequences[7–9]. Finally, institutional, political or financial constraints may limit how much machine learning can improve on existing practices.

This analysis seeks to extend the nascent prediction policy literature in two main ways. First, we evaluate how much machine learning could improve the efficiency of environmental monitoring by allocating inspections to facilities at higher risk of non-compliance, using the pre-eminent water regulation in the United States—the Clean Water Act (CWA)—as our case. Second, we aim to address real-world challenges of applying machine-learning techniques by accounting for institutional constraints and potential strategic behaviour.

Several characteristics of the CWA render it a useful application for using machine learning to enhance public resource allocation. First, similar to many other environmental regulations, limited financial, human and technical resources constrain monitoring and enforcement of the CWA. A diverse set of over 300,000 facilities are regulated under the CWA, ranging from wastewater treatment plants to industrial factories, agricultural operations, shopping centres, and more. Each year, less than 10% of these facilities (~27,500) are inspected, and 25% were inspected at least once between 2012 and 2016. Second, the federal agency overseeing CWA enforcement—the Environmental Protection Agency (EPA)—has a goal of 'innovative and efficient targeting of resources on the most important non-compliance and environmental problems'[10], which aligns well with broader enthusiasm for using 'big data' and inspection targeting in environmental problem solving[11–13]. Its 2017 inspection manual encourages the use of facility records to identify the facilities most in need of inspection[14]. Third, the EPA is collecting increasing amounts of data on regulated facilities, supported by a shift to electronic reporting systems[15]. Machine-learning methods are ideally suited for the task of targeting limited inspection resources: they examine the outcomes of inspections that have already occurred to predict outcomes for facilities that have not been inspected, synthesizing the available data to produce an evidence-based relationship between facility characteristics and inspection failures. Such approaches can support inspectors facing potentially overwhelming amounts of data and can supplement existing protocols to manage pollution.

US state and federal authorities spend US$20–40 million annually on inspections (Supplementary Note 1). These inspections are critical to monitoring and enforcement efforts, and ultimately protecting human health, because they detect many different types of violations, such as substandard management practices, insufficient toxic water treatment protocols and failures to keep accurate discharge records. They have also played an important role in deterring violations, even among uninspected facilities[16].

Determining how to best allocate environmental inspections requires consideration for where to direct inspections, how inspections might increase compliance and the societal benefits of returning to compliance. In this analysis, we address the first component in three main steps.

First, to predict the likelihood of an 'inspection failure' (that is, when an inspector detects and reports a violation at the facility), we train five machine-learning models using publicly available data related to a facility's location, industry and inspection history during 2012–2016 (Supplementary Table 1). We assess the performance of each model by comparing each facility's predicted likelihood of

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

inspection failure (henceforth, its 'risk score') against its observed outcome and select our preferred algorithm based on model prediction accuracy (see Materials and Methods).

Next, we use our best-performing machine-learning model to predict risk scores among a random sample of 20% of uninspected (n=47,181) and inspected (n=15,801) facilities. Using the risk scores, we then evaluate several proposals for reallocating 20% of average annual inspections (n=5,480):

(1) Aggressive, national: inspections are allocated to the highest-risk facilities across the entire United States, based on a rank-ordering of risk scores.
(2) Aggressive, state: inspections are allocated to the highest-risk facilities, but cannot be reassigned from one state to another. This approach preserves each state's observed 'inspection budget' while targeting the highest-risk facilities within each state.
(3) Deterrence, national: maintain a 1% inspection probability for all facilities and assign the remaining inspections based on risk score.
(4) Majors, deterrence, state: inspect 50% of larger, 'major' facilities, maintain a 1% inspection threat for all others, then prioritize the remaining inspections by rank-ordered risk, while preserving state inspection budgets. This allocation aims to match an EPA goal to inspect all major facilities every two years, which we interpret as half of the facilities per year.

To assess the benefits of each reallocation regime, we calculate the predicted additional number of inspection failures relative to the observed inspection failure rate (6.7%), which we refer to as the base rate.

Finally, we consider 'manipulable' data: self-reported pollutant discharge information from facilities that may improve predictions but also exposes the predictive algorithm to manipulation by facilities. These discharge monitoring reports (DMRs) do not substitute for inspections because they include only a small amount of information, but they may help target inspections. Using data from the subset of facilities that submit these reports, we test whether the self-reported data improve predictive power by developing and evaluating two models: one with the original set of covariates and one with additional covariates from the DMRs.

Results

Model performance. Among the five algorithms we tested, the regression forest performed the best based on the area under the receiver operating characteristic curve (AUC) criterion (AUC=0.922; Supplementary Table 3). While all five models were highly accurate, the regression forest correctly predicted inspection outcomes for 94.1% of facilities. Additionally, of the predicted passed inspections, 95% were actually passed, which indicates that few violations are going undetected. Comparing the rate of inspection failure across the distribution of predicted risk scores shows that inspection failure increases from less than 1% among facilities with risk scores between 0 and 0.05 to over 80% for facilities with risk scores between 0.80 and 0.85 (Fig. 1). The distribution of risk scores is shown in Supplementary Fig. 4.

Inspection reallocation. The data-driven inspection allocations increase the rate of failed inspections from the 6.7% base rate to a predicted 13–50%, depending on allocation constraints (Fig. 2 and Supplementary Table 5). As shown in Fig. 3, the current allocation inspects many low-risk facilities, and there is substantial room for improvement. Under the least-constrained reallocation, in which inspections are allocated nationwide to the facilities with the greatest predicted risk of failing an inspection, the percentage of inspections that detect violations would increase from about 6.7 to 50%—an increase of more than 600%.

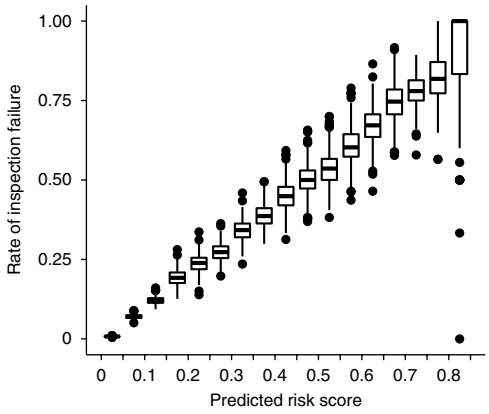

Fig. 1 | Rate of inspection failure by predicted risk score. The likelihood of inspection failure increases with the predicted risk score. The box plots show the distribution of observed inspection failure rates for risk score bins. The centre line within each box marks the median, and the limits of each box represent the 25th and 75th percentiles. Whiskers indicate 1.5× the interquartile range, and points indicate outliers beyond 1.5× the interquartile range. Distributions are generated by evaluating 500 regression forest model runs, where we randomly select a different 80–20% training and test partition and calculate inspection failure rates among the 20% test set (n=15,801) from each run.

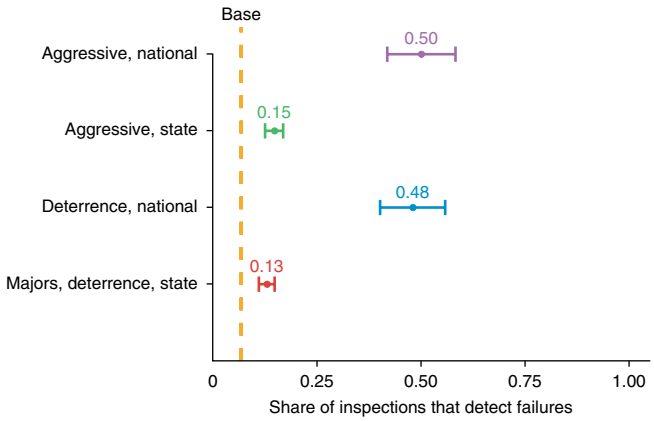

Fig. 2 | Inspection failure rates under different reallocations. The risk-based reallocations significantly increase the predicted rate of violations detected from inspections relative to the base case. The vertical dashed line represents the base case, which is the current share of inspections that detect violations. The points indicate the average share of inspections that detect violations per reallocation proposal. Error bars represent the 95% confidence interval (see Supplementary Table 5 for numerical values). Risk-based inspection allocation could increase the detected failure rate from 6.7 to 13–50% of inspections.

As we impose additional constraints, more inspections are assigned to lower-risk facilities, which reduces the predicted number of violations detected. Under the most constrained reallocation—which preserves state-level inspection levels, maintains a 1% probability of inspection for all facilities and inspects 50% of major facilities each year—we find that the expected rate of failed inspections doubles compared with current practices (from 6.7 to 13%).

Value of self-reported data. For the subset of facilities that submit DMRs to the EPA (n=~43,000), the inspection failure rate is 8.1%—slightly more than the 6.7% inspection failure rate among our broader facility dataset. We find no substantial difference between the

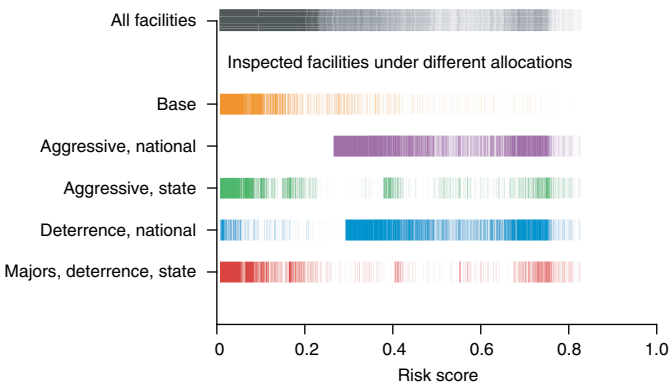

Fig. 3 | Inspected facilities under different allocations. The proposed reallocations direct inspections to higher-risk facilities. Each vertical line in the strip plots represents a facility. The top strip plot indicates where all facilities in our final analysis set fall along the risk score distribution ($n = 62{,}982$). In the 'base' strip plot, only facilities that are inspected under the current allocation are included. The four strip plots below this depict only the facilities assigned to inspection under each reallocation proposal ($n = 5{,}480$ inspections).

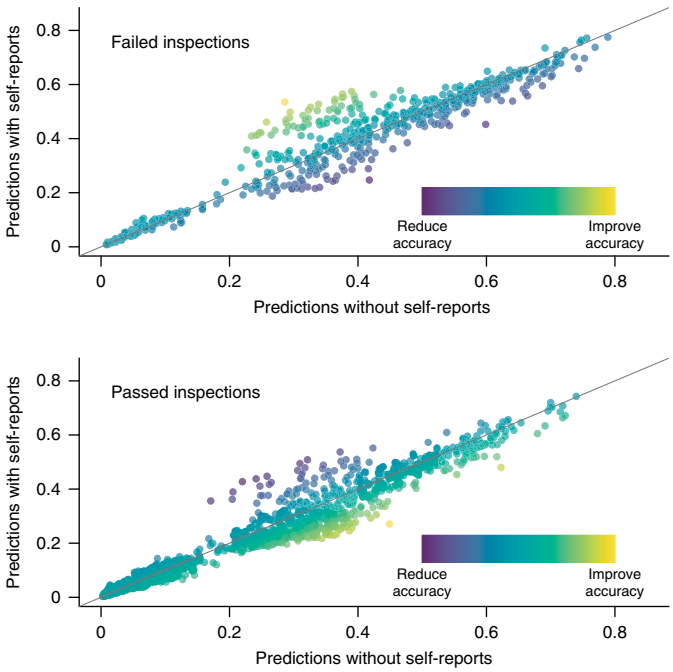

Fig. 4 | Influence of self-reported data on risk scores. Including the self-reported DMR information improves the accuracy for some facilities and reduces it for others. Predicted risk scores for each facility are shown with (*y* axis) and without (*x* axis) the data from DMRs. The closer the scores are to the 45° lines, the less the additional information changed the predicted risk of failing an inspection. The colour gradient indicates the extent to which adding in self-reported information from the DMRs improves the accuracy of the predictions, with lighter, yellow shading indicating an improvement and darker, purple shading indicating a reduction in accuracy.

models trained with and without the additional data from the DMRs. In addition, including the self-reported information did not uniformly improve prediction accuracy; rather, accuracy increased for some facilities and declined for others. Figure 4 compares and contrasts the two models' predicted risk scores. Additional detail on performance with and without the self-reported data is shown in Supplementary Table 4.

Discussion

In this paper, we demonstrate how machine-learning techniques can enhance the ability of public agencies to meet their regulatory objectives. We train a regression forest model to predict which facilities are likely to fail an inspection and use these predicted risk scores to suggest alternative allocations of CWA inspections. Compared with current practice, our alternative reallocation schemes nearly double the rate of violations detected through inspections in our most constrained reallocation, and increase it by over 600% in our most aggressive reallocation.

The reallocation scenarios we propose are designed to consider several real-world constraints on the EPA. First, we consider a geographic constraint: state-level inspection budgets. Of all the restrictions we impose, state-level inspection budgets reduce the predicted number of violations detected through inspections the most. This result is due to substantial differences in the number of inspections conducted per state compared with the locations of high-risk facilities. If it were possible to pool inspection resources across the country, inspections could be reallocated from Pennsylvania, which has many inspections but few high-risk facilities, to states such as Montana, Kentucky or Texas, each with many high-risk facilities.

Second, we address strategic responses by regulated facilities by evaluating a reallocation alternative that ensures a minimum probability of inspection across all facilities. With a nationwide reallocation of inspections, imposing a 1% probability of inspection for all facilities slightly reduces the predicted benefits, but the number of inspection failures is still predicted to increase by about 600%.

Third, we add a constraint from 2017 CWA inspection guidelines: inspecting 50% of the larger, 'major' facilities each year. Combined with the state-level budgets and minimum probability of inspection, this is the most highly constrained reallocation. Even while meeting all of these constraints, this reallocation detects nearly double the failures of the base allocation.

Our analysis also addresses the trade-offs associated with incorporating self-reported, manipulable data that may improve predictions. In this case, we find that publicly available self-reported data from DMRs do not substantially improve the predictive power of the model. The finding that self-reported effluent records do not particularly improve predictive accuracy may be explained by at least two points. First, inspections record more than effluent levels, and violations may be identified in management or measurement practices. Second, facilities that self-report effluent exceedances may respond on their own to return to compliance, weakening or eliminating the link between a self-reported effluent exceedance and an inspection failure.

On the one hand, since including self-reported data does not improve the predictive power of our models, these data could be omitted without significant consequences for the reallocation of inspections. Excluding self-reported data from the model can help reduce concerns about strategic behaviour and manipulation by facilities responding to an algorithm-based inspection allocation, as well as the limited ability of the EPA to detect fraudulent reports[12]. On the other hand, regulators could include this information, recognizing that these potentially manipulated data do not dramatically change risk scores (and thus the probability of inspection) in the current model.

We acknowledge several limitations to our machine-learning-based application. First, this approach—as with any that makes out-of-sample predictions—is most appropriate when the observed population sample offers transferable knowledge about the population at large[17,18]. We model the risk of failing an inspection based on observed (inspected) facilities and apply that same model to unobserved (uninspected) facilities. While the subset of inspected facilities is not random, we find substantial overlap in terms of risk scores (Supplementary Fig. 4), which also indicates overlap between the characteristics of inspected and at least some uninspected facilities.

Second, our approach does not account for potential changes over time. We train our models on data observed from the last five years (2012–2016), since we suspect that public policy priorities, pollution control technologies and the types of inspection failures were relatively constant during that time frame. Our predictions therefore apply to inspection failure during this same time period. However, the factors that are most predictive of inspection failure may change over time. Furthermore, facilities may strategically respond after an algorithm-based inspection regime is adopted. For example, data may be manipulated to produce a lower risk score, or if managers perceive their likelihood of inspection to be low, they may relax their performance standards. How much each of these strategic responses is a concern depends on how much a facility can change the factors linked to its risk score, the number of inspections that are randomly allocated and the penalties associated with data manipulation, among many other factors. A real-world application should consider how to monitor and account for such changes over time.

Third, an important factor to consider with all data-driven approaches is that they may codify or exacerbate existing biases and forms of discrimination[7,19]. In the case of the CWA, this could potentially exacerbate environmental justice concerns if a data-driven approach systematically directs oversight away from facilities located in low-income or minority areas. In our application, the predicted risk scores were responsive to variables related to environmental justice concerns, as facilities that fell within the 80th percentile of the environmental justice screen index also tended to have higher risk scores (Supplementary Fig. 6). A real-world implementation could be more robust to environmental justice concerns by building them directly into inspection targeting. For example, similar to the requirement to regularly inspect all major facilities, an alternative allocation protocol could require inspections for a specified share of all facilities that fall above the 80th percentile of the EPA's environmental justice screen index.

This analysis demonstrates the potential benefits of adopting a more data-driven inspection allocation regime. Our model serves as a starting point that could be augmented with greater detail on the costs and benefits of different inspections, violations and enforcement responses. Future work could examine additional complexities of integrating a machine-learning approach into the EPA's broader enforcement efforts, such as incorporating specific enforcement priorities or identifying technical, financial and human resource limitations[12]. In addition, these methods could be applied in other contexts within the United States and beyond, where regulators are seeking to make efficient use of limited resources. We are optimistic that researchers and public agencies can continue to expand and test these methods to improve environmental sustainability.

Methods

Data. We primarily used two data sources from the EPA: the Integrated Compliance Information System (ICIS) and facility-level data from the Enforcement and Compliance History Online (ECHO) database. Specifically, we used data on all facilities recorded in the EPA ECHO database as of 16 April 2017. The original data file included 1,831,032 facilities, of which 393,220 had permits under the National Pollutant Discharge Elimination System (NPDES) and were thus relevant for analysis. Among facilities with active permits, we focused on 316,030 facilities with complete information on facility characteristics.

Inspection events and violations were identified using ICIS data, which include both state and federal inspections. Each inspection event was marked with a permit number and date. To identify inspections that resulted in violations, we matched inspection dates to NPDES single-event violation records. As a conservative measure, we defined inspection failures as when inspections and single-event violations for the same permit were recorded for the same day. Other violation types in the ICIS data (for example, permit schedule violations) were not considered here because they were generated by the system automatically. Single-event violations had to be entered manually and therefore contained the violations identified through inspections[20]. Based on this definition of inspection failure, 6.7% of inspections in our dataset were failed.

While single-event violations and inspections were recorded at the permit level, a single facility could hold multiple permits. As our unit of interest was the facility, we aggregated inspections and violations for separate permits up to the facility level. For example, if a facility had three permits and one was flagged for a single-event violation, we considered the facility to have failed its inspection.

ECHO provides facility information such as location, industry and nearby population density. We also constructed several additional variables that may affect compliance likelihoods, such as the number of days since the last inspection, number of CWA-regulated facilities in the county and in the state, number of facilities within 50 miles inspected in the year, two years and five years before inspection, and 2017 political affiliation of the governor of the state or territory. Supplementary Table 1 presents a full description of all the variables included in our analysis and indicates which were constructed.

In addition, we evaluated the additional predictive power of including data from DMRs, which are periodic self-reports of pollutant levels submitted to the EPA that are typically required of facilities that discharge wastewater from a pipe. To compare the predictive accuracy with and without self-reported information, we restricted our sample to facilities that submitted DMRs and repeated the model development procedure described above twice: once with the original set of covariates available for all facilities and once with four additional covariates we constructed from the DMRs (the number of DMRs submitted, number of late DMRs (D80 and D90 violations) and number of DMRs that report having exceeded pollutant limits (E90 violations)), all calculated in the year before inspection.

For facilities that were not inspected, we constructed three values to permit comparison. First, for days since the last inspection, we used the median value among inspected facilities. Second, when we calculated the number of proximate facilities inspected in the year, two years and five years before inspection, we assumed a hypothetical inspection date of 1 January 2017. Finally, we assumed the same hypothetical inspection date when we calculated the DMRs submitted and D80, D90 and E90 violations in the year before inspection.

For facilities inspected multiple times from 2012 to 2016, we randomly selected only one inspection per facility to include in our analysis. These inspections were combined with a 20% sample of uninspected facilities for the final dataset. Of the inspected facilities in our final dataset, 58.5% of facilities were inspected once over 5 years, and 41.5% were inspected more than once. We chose to select only one inspection per facility because some inspections were follow-up inspections to check whether or not facilities had returned to compliance. The current approach preserves this follow-up inspection budget and avoids overweighting facilities that are inspected multiple times.

Predictive models. To predict the risk of failing an inspection, we tested five different algorithms: logistic regression, least absolute shrinkage and selection operator (LASSO), elastic net, regression tree and regression forest in R version 3.4.2 (ref. [21]). These five algorithms were chosen due to their: (1) common use in machine-learning classification/prediction problems; (2) demonstrated performance and predictive power; and (3) differing abilities to address overfitting. Logistic regression was included to serve as a comparison of a more standard regression approach to the machine-learning approaches. Briefly, we describe each of the other four algorithms used:

(1) The LASSO is a form of regularized regression that minimizes the sum of squared residuals subject to the sum of absolute values of the model parameters being less than some specified constant. Imposing this restriction shrinks some coefficient values to zero. The constant, known as a tuning parameter, controls the extent of the shrinkage. The optimal tuning parameter should be selected using cross-validation. By shrinking some coefficients to zero, the LASSO estimator trades off some bias for reduced variance, thus improving the predictive power of the model over a standard ordinary least square (OLS) model. Additionally, because coefficients are shrunk, the LASSO can be used for model selection.

(2) The elastic net is a generalized extension of the LASSO that improves performance when the number of covariates is greater than the sample size or the covariates are highly correlated. The elastic net imposes two penalty terms: (1) the sum of the absolute value of the model parameters; and (2) the sum of squared model parameters in a linear combination. In this way, the elastic net operates as a convex combination of the LASSO (which uses only an absolute value penalty term) and ridge regularized regression (which uses only a squared penalty term) estimators.

(3) Single regression trees are a form of decision tree learning that can be used as a simple prediction method. The data are split into partitions (leaves) based on covariate values and then predictions are made using the average value of the outcome within each leaf. To prevent overfitting, trees are 'pruned' by selecting a tree size that minimizes the cross-validated error.

(4) Regression forest is a generalization of the single regression tree approach, where many trees are built and the predictions are averaged. Regression forests improve on the single regression tree approach without overfitting, so individual trees do not need to be pruned, since many trees are averaged.

The key to this approach is that the prediction is a weighted average, where predictions are weighted by the number of trees in which a given observation ends up in the same leaf as each of its neighbouring observations. This weighted-average approach sets the regression forest apart from the commonly used random forest.

To avoid overfitting, we employed standard cross-validation practices. First, we randomly split the inspected facilities into training (80%) and test (20%) samples. We then modelled the probability of failing an inspection in the training data as a function of the facility characteristics. One advantage of machine-learning approaches is that they identify which variables are most useful for making predictions. Thus, we included all available data related to those facilities (as featured in Supplementary Table 1).

We created two-way interactions between all variables for the LASSO and elastic net models. The regression tree and regression forest non-parametrically search across all covariates and interactions, eliminating the need to manually include interaction terms for them. The logistic regression included only the main variables and no interactions due to concerns of overfitting and loss of degrees of freedom with high numbers of predictors.

To select tuning parameters for the LASSO and elastic net, we used tenfold cross-validation. After training each model, we predicted risk scores on the test data and assessed the accuracy of each model's predictions by comparing the predicted risk score with the observed inspection outcomes in the test data. We assessed the performance of the models using the AUC criterion, which combines the true positive and false positive rates in a single metric of overall predictive performance. We selected the model with the highest AUC for the remainder of the analysis. We also evaluated model performance by comparing the mean squared error, sensitivity (true positive rate), specificity (true negative rate) and accuracy (that is, the rate of correct predictions out of all predictions) (Supplementary Fig. 3 and Supplementary Table 3). To confirm that our preferred model is robust to different partitions of the data into training and test sets, we implemented a fivefold cross-validation approach, which involved dividing the data into five randomly selected datasets (or folds), and using four of the folds to train the model and the fifth fold as a test set. This was repeated until each fold had been used as both training and test data. The regression forest result was robust to this cross-validation and thus utilized for the remainder of the analysis.

Two different sets of facilities were considered: first, all facilities, and second, only facilities that submitted DMRs. Using the ~43,000 DMR-submitting facilities, we developed two models. The first model used the same predictors as the original model. Then, we used the same facilities but included the additional information provided by the DMRs. We implemented the same approach for predicting risk scores, assessing the performance of the models and testing the sensitivity of the result to different partitions of the data into test/training sets among the facilities that submitted DMRs twice: first with the same predictors as the original model and then again with inclusion of the DMR data. The regression forest performed the best in both cases.

Failure rates. To estimate the relationship between predicted risk scores and inspection failures, we divided our test data based on 0.05 intervals of risk scores (that is, facilities with risk scores between 0 and 0.05 were in one group, those with risk scores between 0.05 and 0.10 were in the next, and so on). We calculated the observed failure rate for each group. We repeated the process with 500 random 80/20 training/test partitions (running regression forest, grouping facilities by risk score and calculating the observed failure rate). We used the rates from the 500 runs to estimate a mean and s.e. for the failure rate of each risk score group.

Reallocation proposals. To evaluate the benefit of using a data-driven approach to target inspections, we proposed four inspection reallocations and compared them with current practice (the base case). We predicted risk scores for a combined dataset of 20% of the inspected facilities (our test set) and 20% of uninspected facilities. Then, we reallocated a proportionate number of inspections (20% of the annual average). Our four reallocation protocols were as follows:

(1) Aggressive, national: allocate all inspections to the highest-risk facilities across the entire United States, based on a rank-ordering of risk scores. For example, if we had 1,000 inspections to assign, the facilities with the 1,000 highest risk scores would be assigned to inspection.
(2) Aggressive, state: allocate inspections to the highest-risk facilities per state but using state inspection budgets. We maintain the same number of inspections per state as their observed average, then distribute inspections to the highest-risk facilities in each state. These 'state inspection budgets' prevent inspections from being reassigned from one state to another.
(3) Deterrence, national: maintain a 1% inspection probability for all facilities and prioritize the remaining inspections by risk score. To implement this, we assign each risk score group a number of inspections equal to 1% of the total number of facilities in the group. These inspections are randomly assigned. Then, with however many inspections remain, we assign inspections to the highest-risk facilities.
(4) Majors, deterrence, state: inspect 50% of larger, 'major' facilities, maintain a 1% inspection threat for all others, then prioritize the remaining inspections by rank-ordered risk, all using state inspection budgets. The EPA directs states to inspect all 'major' facilities every two years[22]. While states can decide when to conduct these inspections, we account for this constraint by requiring 50% of major facilities to be inspected. Second, we maintain a deterrence effect by randomly assigning inspections to 1% of facilities. Any inspections left over are directed to the highest-risk facilities in that state until inspection budgets are exhausted.

For each reallocation proposal, we 'assigned' inspections to facilities according to the rules of each reallocation regime, using a fixed inspection budget, and multiplied the failure rates by the number of facilities in each 0.05 score interval grouping. More formally, where the risk score group was represented by i, the total violations detected per allocation was calculated as:

$$\text{violations} = \sum_{i=1}^{n} (\text{mean failure rate}_i \times \text{number of facilities}_i) \quad (1)$$

As an additional evaluation, we repeated the reallocations among only inspected facilities, as the outcome for those facilities was known and therefore did not require prediction. The results of this analysis are shown in Supplementary Table 6. Finally, to demonstrate how the benefits of machine-learning approaches vary depending on resource constraints, we repeated our reallocation procedure with varying numbers of inspections to allocate. As shown in Supplementary Fig. 5, the benefits decrease as the number of inspections increase because the risk score is less influential if 80, 90 or 100% of facilities are going to be inspected.

Data availability

The raw data used in this analysis can be downloaded from the EPA's ECHO website (https://echo.epa.gov/). The processed datasets are also available with code at the Stanford Digital Repository (https://purl.stanford.edu/hr919hp5420).

Received: 2 March 2018; Accepted: 23 August 2018;
Published online: 01 October 2018

References

1. Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. Prediction policy problems. *Am. Econ. Rev.* **105**, 491–495 (2015).
2. Athey, S. Beyond prediction: using big data for policy problems. *Science* **355**, 483–485 (2017).
3. Mullainathan, S. & Spiess, J. Machine learning: an applied econometric approach. *J. Econ. Pers.* **31**, 87–106 (2017).
4. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. Human decision and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018).
5. Kang, J. S., Kuznetsova, P., Luca, M. & Choi, Y. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In *Proc. 2013 Conference on Empirical Methods in Natural Language Processing* 1443–1448 (Association for Computational Linguistics, 2013).
6. Chandler, D., Levitt, S. D. & List, J. A. Predicting and preventing shootings among at-risk youth. *Am. Econ. Rev.* **101**, 288–292 (2011).
7. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, New York, USA, 2016).
8. Blumenthal-Barby, J. S. & Krieger, H. Cognitive biases and heuristics in medical decision making. *Med. Decis. Making* **35**, 539–557 (2015).
9. Mullainathan, S. & Obermeyer, Z. Does machine learning automate moral hazard and error? *Am. Econ. Rev.* **107**, 476–480 (2017).
10. Lund, L. C. *Clean Water Act National Pollutant Discharge Elimination System Compliance Monitoring Strategy* (United States Environmental Protection Agency, 2014); https://www.epa.gov/sites/production/files/2013-09/documents/npdescms.pdf
11. Friesen, L. Targeting enforcement to improve compliance with environmental regulations. *J. Environ. Econ. Manage.* **46**, 72–85 (2003).
12. Rivers, L., Dempsey, T., Mitchell, J. & Gibbs, C. Environmental regulation and enforcement: structures, processes and the use of data for fraud detection. *J. Environ. Assess. Pol. Manage.* **17**, 1550033 (2015).
13. Glicksman, R. L., Markell, D. L. & Monteleoni, C. Technological innovation, data analytics, and environmental enforcement. *Ecol. Law. Q.* **44**, 41–88 (2017).
14. *NPDES Compliance Inspection Manual* Interim Revised Version, January 2017 (United States Environmental Protection Agency, 2017); https://www.epa.gov/sites/production/files/2017-01/documents/npdesinspect.pdf
15. *National Pollutant Discharge Elimination System (NPDES) Electronic Reporting Rule* (United States Environmental Protection Agency, 2015); https://www.gpo.gov/fdsys/pkg/FR-2015-10-22/pdf/2015-24954.pdf
16. Shimshack, J. P. & Ward, M. B. Enforcement and over-compliance. *J. Environ. Econ. Manage.* **55**, 90–105 (2008).
17. James, G., Witten, D., Hastie, T., & Tibshirani, R. *An Introduction to Statistical Learning* (Springer, New York, USA, 2013).

18. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* 2nd edn (Springer, New York, USA, 2009).
19. Zliobaite, I. Fairness-aware machine learning: a perspective. Preprint at https://arxiv.org/abs/1708.00754 (2017).
20. *ICIS-NPDES Download Summary and Data Element Dictionary* (United States Environmental Protection Agency, 2017); https://echo.epa.gov/tools/data-downloads/icis-npdes-download-summary
21. R Development Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
22. *State Compliance Monitoring Expectations* (United States Environmental Protection Agency, 2015); https://echo.epa.gov/trends/comparative-maps-dashboards/state-compliance-monitoring-expectations

Acknowledgements

We thank S. Athey, M. Burke, F. Burlig, K. Mach, A. D'Agostino, C. Anderson, K. Green, S. Hasan, D. Jiménez, H. Kim, A. R. Siders and A. Stock for comments. E.B. receives funding from the National Science Foundation Graduate Research Fellowship Program (DGE-114747), M.H. from the Department of Earth System Science at Stanford University, and N.B. from the Stanford Graduate Fellowship/David and Lucile Packard Foundation.

Author contributions

All three authors collaboratively designed the study, developed the methodology, assembled the data, wrote the code, performed the analysis, interpreted the results, and wrote the manuscript. E.B. and M.H. conducted the final analysis, with substantial input from N.B.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41893-018-0142-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.