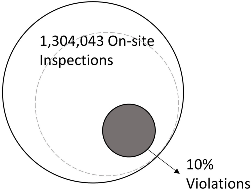
Better targeting through machine learning in environmental enforcement

Jiaqi Zhang, Mengdi Liu, Bing Zhang
dg21250055@smail.nju.edu.cn

CEMPA
Nanjing University

Challenge of Environmental Enforcement

Environmental agencies spend millions of dollars annually on pollution control, but fail to detect and penalize most violators (Hino et al., 2018). Why?



Environmental inspection in China (2020)

- Limited inspection resources: regulators' incomplete grasp of information about polluters (Andarge, 2019); the growing number of facilities (Shimshack, 2014).
- Inefficient enforcement resource allocation: regulators' reliance on private information based on few firm characteristics (Benjamin, 2018; Blundell et al., 2020; Earnhart & Friesen, 2021); imperfect human decision making (Kleinberg et al., 2015).

Use machine learning to enhance enforcement

- With the rapid development of monitoring and big data technologies, machine learning has become a beneficial tool for public agencies (Athey, 2017).
  - Accurately and objectively assess the situation (Kleinberg et al., 2015).
  - Rationalize the allocation of limited resources and maximize the value of limited resources through data prediction (Athey, 2017).
- Machine learning can also be used to predict a facility's risk (Hino et al., 2018).
- This research
  - Achieve better targeting in environmental enforcement through machine learning, using the example of environmental regulation in Jiangsu Province, China.
  - Increase the rate of violation detection by 40.43% to 148.16% per month.
  - Cut 61,786 inspections in Jiangsu Province in 2020, saving 57 million CNY.

Contributions

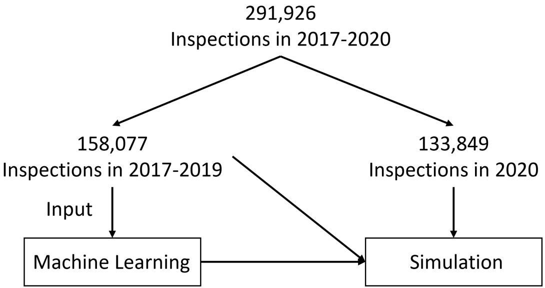
- Extend the literature on enforcement resource optimization (Hino et al., 2018; Gerarden & Yang, 2023).
  - Use a novel GBM-based model to predict the environmental violation probability of facilities
  - Use it to optimize the inspection planning under limited resources.
  - Explain the model behind the black box model - inspection interval and violation status.
- Extend research on environmental enforcement in China (Zhang et al., 2018; Xiao et al., 2023; Wang et al., 2023).
  - Introduce machine learning into monthly inspection plan in China.
  - The machine learning-based inspection is more effective than the current inspection.
- Extend research on the application of machine learning methods (Hino et al., 2018; Chang et al., 2020).
  - The superiority of the GBM model in predicting the probability of environmental violations. GBM>Random Forest>Linear Models.

Data

- Our study is based on data in Jiangsu province between 2017 and 2020.
- We used various data from authoritative databases provided by the Department of Ecology and Environment.
    - Enforcement features: Administrative Penalty Database and Environmental Inspection Database
    - Facility characteristics: Discharge Permit System and Business Database
    - Online monitoring information: Continuous Emission Monitoring System (CEMS) Database
    - Degree of being regulated: Key Emission Unit List
    - Public Complaints: "12369" Complaints Database
    - Environmental quality: Real-time air quality data at station level, and PM2.5 dataset derived from satellite remote sensing data (Shao et al., 2020)

Data

- After data matching, we focused on enforcement records with complete facility characteristic information and investigation results from 2017 to 2020, totaling 291,926 records.



Method: Outcome Variable

- Violation - the investigation result recorded
  - The investigation result is stored in the inspection database as a binary variable, which is 1 if a violation was detected in this inspection; otherwise, it is 0.
  - Manually recorded by the regulator after the on-site inspection, and is an official investigation result with reliability.

Method: Predictive Variables

$P\,(Violation)_{i,t} = f\,(Inspection_{i,t} + Penalty_{i,t} + Facility_i + Key_{i,t-1tot} + CEMS_{i,t-1tot} + Resour_{i,t-1} + Complaint_{i,t-1tot} + AirQulity_{i,t-1tot})$

- Enforcement features:
  We first construct several variables about inspections and penalties that may affect the probability of violation, such as
  - The number of days since the last inspection
  - The investigation results of the last inspection
  - The number of inspections recorded in the database for the same facility
  - Environmental penalties in the last month.

Method: Predictive Variables

- Facility characteristics:
  - Basic information about the facility, such as registered capital, number of insured persons, industry and the type of ownership of the facility (SOE and FOE).
  - Additional environmental information about the facility, such as the major pollutant categories recorded in the permit, the number of fixed outfalls.
- Key Emission Unit
- CEMS monitoring data
  - To capture whether the facility is being monitored in real-time.
- The inspection resources divided by the number of facilities in the city from the discharge permit database
  - Imply the probability of being inspected per facility in the city.
- Public complaints
- Environmental quality

Method: Predictive models

- In order to predict the probability of finding a violation in a single inspection event, we tested five different algorithms:
    - Logit
    - Penalized Logit (from `caret` pkg): including Ridge Regression, LASSO and Elastic Net
    - CART (from `caret` pkg): Classification and Regression Tree
    - Bagging (from `caret` pkg): Bagging Tree
    - RF (from `caret` pkg)： Random Forest
    - GBM (from `caret` pkg): Gradient boosting trees
- The five algorithms were chosen because of
    - Tree-based machine learning models have been shown to be superior in predicting the environmental risk of facilities compared to linear parametric regression models such as Logit and LASSO.

Method: Model evaluation

- Evaluate the performance of the trained model on the test set
  - Randomly split: training set (80%) & test set (20%).
  - Methods for evaluating predictive power: Probability Calibration Plot; Accuracy and Precision calculated by confusion matrix; Receiver Operating Characteristic (ROC).

Method: Feature analysis

- Investigate the impact of different types of environmental variables on the predicted violation probabilities of facilities to improve the transparency and reliability.
- Feature Importance: demonstrate how important the feature is in the prediction.
- Partial Dependency Plots: depict the pattern of output variables changing with the feature of interest, with other features remaining constant.

Method: Model application in monthly inspction plan

Updated Monthly Facility List for Simulation

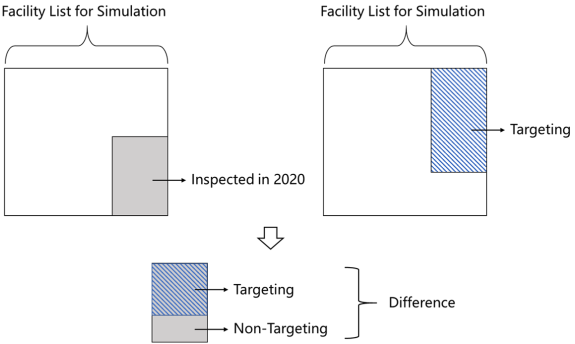
Input: Facilities from records before 2020.01 ⇨ Final Predictive Model ⇨ Output: $P_{i,1}$(violation) in 2020.01

Input: Facilities from records before 2020.02 ⇨ Final Predictive Model ⇨ Output: $P_{i,2}$(violation) in 2020.02

…… ……

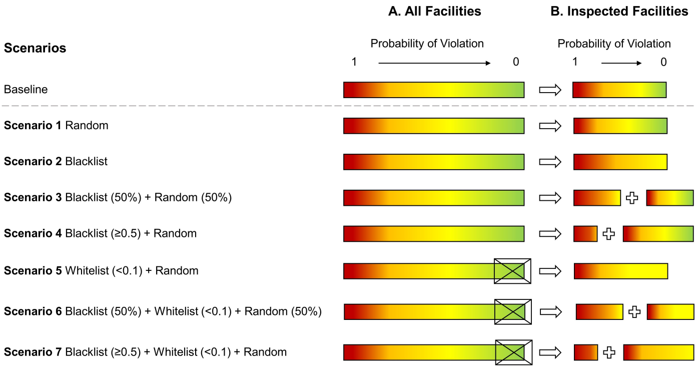
Input: Facilities from records before 2020.12 ⇨ Final Predictive Model ⇨ Output: $P_{i,12}$(violation) in 2020.12

For each month:

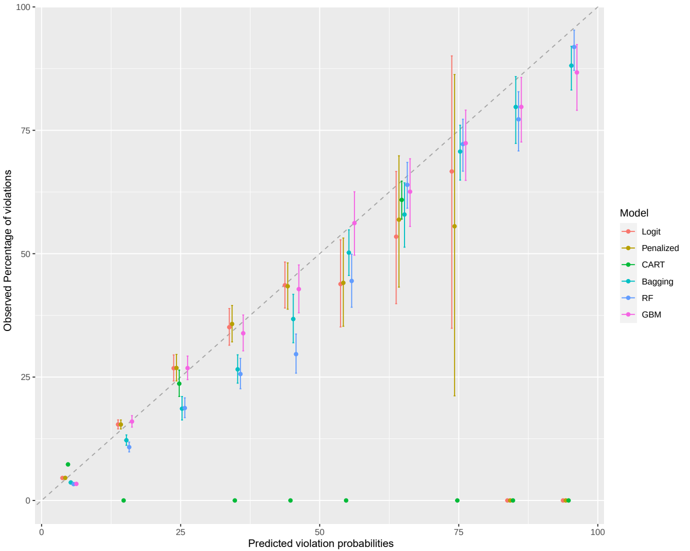


Method: Model application in the context of introducing a randomly selected inspection scenario

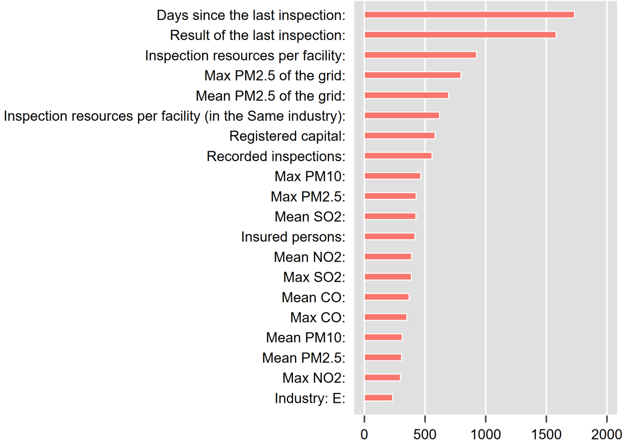
- Models may over-regulate certain firms or industries based on historical data bias.
- Propose seven inspection scenarios combining random inspections to reduce the drawbacks of algorithm-based resource allocation.



Predict the probability of violation: Calibration Curve

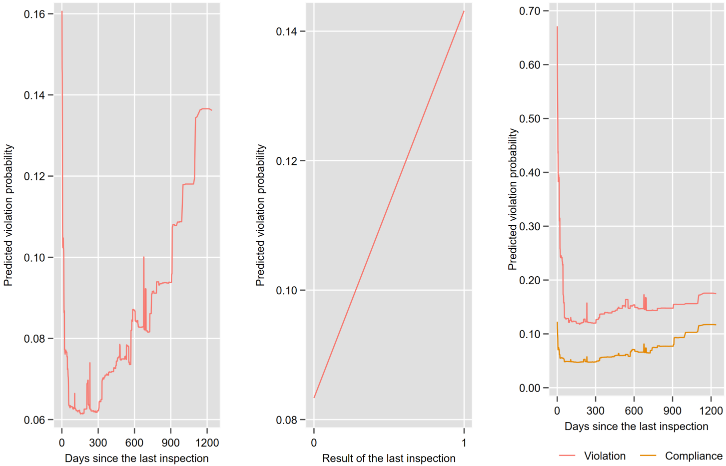


Feature Importance

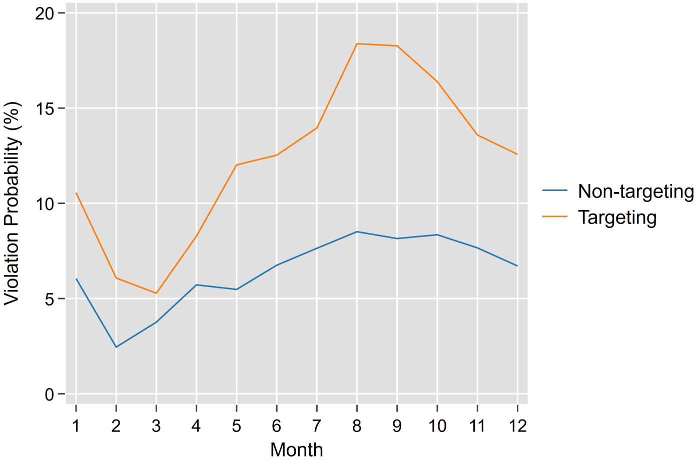


Among these 20 variables, the two most important variables are those related to environmental enforcement (Shapiro & Walker, 2018; Telle, 2009, 2013).

Interaction between features

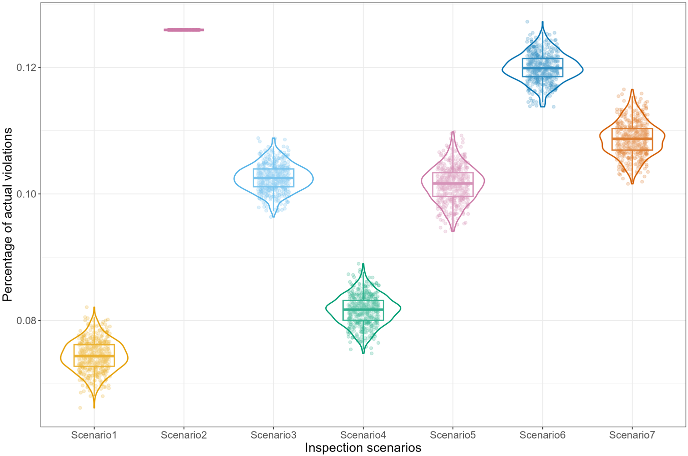


Increase the accuracy of targeting inspections in 2020



Increase the rate of violation detection by 40.43% to 148.16% per month; Cut 61,786 inspections in Jiangsu Province in 2020, saving 57 million CNY.

Combine random and targeting inspections



Scenario 6 performs the best among the seven examination options.

Conclusions

- We demonstrate the ability of machine learning techniques to assist environmental inspectors in achieving better targeting.
- Next step:
  - Explore the patterns of interaction between polluting facilities and enforcement agencies implied behind the model.
  - Measure whether machine learning models reinforce bias in the data and whether introducing random sampling reduces such bias.

Thanks!
Comments are welcome!