LAW INSPECTIONS

Using big data to accomplish a huge job

Are we making the best use of data in our efforts to protect the natural environment? A new study suggests that we could be doing much better.

Louie Rivers III

Over the past 40 years, the budget for the Environmental Protection Agency (EPA) in the US has been reduced by 55% when adjusting for inflation¹. The current administration's most recent budget proposal calls for a further 23% reduction of the agency's budget². This is despite the agency's responsibilities drastically expanding over the same time period, the population of the United States increasing by 100 million people and considerable economic growth¹. It is clear that for the foreseeable future, the protection of the environment will require regulators to do more with less. At the same time, the EPA, like many federal agencies, is awash with data that is currently underutilized.

The need for more-efficient use of resources in protecting the environment is at the heart of an Article by Elinor Benami and colleagues³, published in this issue. They build on previous advances in the use of big data and machine learning (ML), using statistical techniques to enable computer systems to learn from data without explicit programming to examine the allocation of scarce environmental-protection resources. The authors apply ML techniques to model the reallocation of public resources to ensure greater compliance with the US Clean Water Act.

Specifically, they make use of real-world data to calculate the optimal allocation of inspections to facilities that are more likely to violate clean-water regulations. The dataset used to develop and test the ML modelling efforts is comprised of the 316,030 facilities that are listed in the National Pollutant Discharge Elimination System with complete facility characteristics listed, and whose regulatory history is featured in the Enforcement and Compliance History Online. The facility characteristics can be categorized as location, industry and inspection compliance history.

Using the Integrated Compliance Information System, Benami and colleagues identified facilities that were inspected during the study period 2012–2016, and determined whether they were in violation of the Clean Water Act. Subsequently,

they tested 5 ML models using facility characteristics to predict a risk score, the likelihood of failing an inspection, for each of the inspected facilities. For each of the ML models, they compared the predicted risk scores with the observed inspection violations and chose the most accurate ML model, known as a regression-forest technique⁴. This model correctly predicted inspection outcomes for 94.1% of inspected facilities.

The research team applied the regression-forest model to predict scores for a sample of inspected and uninspected facilities, and used these risk scores to reallocate inspections using four different protocols. The “Aggressive, national” protocol allocates inspections to facilities with the highest risk scores across the US, regardless of state boundaries. The “Aggressive, state-level” protocol allocates inspections to facilities with the highest risk scores by state. The “Deterrence, national” protocol allows for a 1% chance of all facilities being inspected and allocates the remaining inspections to facilities with the highest risk scores regardless of state boundaries. The “Majors, deterrence, state-level” protocol randomly inspects 50% of the larger facilities, allows for a 1% chance of inspection of all other facilities and allocates the remaining inspections to facilities with the highest risk score by state.

Each of the protocols led to a greater percentage of inspections detecting a violation than do current procedures, which detect a violation for 6.7% of inspections. The two national-level protocols, “Aggressive, national” and “Deterrence, national”, were the most successful at identifying violations, at 50% and 48% respectively. The two state-level protocols, “Aggressive, state level” and “Majors, deterrence, state-level”, were less successful than the national protocols but still approximately doubled the number of inspections leading to violations compared to current procedure, with 14% and 13% of inspections respectively detecting a violation.

The work of Benami and colleagues offers two main innovations for environmental regulation. The first innovation is the use

of big data to improve environmental-inspection practices. Environmental inspectors at the state and federal level are overwhelmed with the amount of facilities that they are required to monitor. Additionally, they are awash with copious amounts of data that are underused at best, and frequently neglected in setting regulatory priorities. The effective marshalling of big data to prioritize environmental regulatory practices is a major innovation.

Benami and colleagues' second contribution builds directly on this innovative use of big data. They use machine-learning techniques and existing data to develop a predictive risk score, the likelihood of failing an inspection, for an incredibly large number of facilities. They then use these predictive risk scores to allocate inspections based on the risk of a facility failing inspection. This is a major change from standard regulatory practices. Currently, significant resources are dedicated to the inspection of low-risk facilities. The researchers suggest an innovative alternative approach; allocating regulatory resources to those facilities that pose the greatest risk to public health.

The work of Benami and coworkers represents a potentially important step in more-efficiently protecting the environment in an era of growing challenges but shrinking budgets. □

Louie Rivers III

North Carolina State University, Raleigh, North Carolina 27695, USA.

Published online: 15 October 2018
<https://doi.org/10.1038/s41893-018-0157-2>

References

1. Gabry, K. EPA's budget has been devastated for decades: Here's the math. *The Hill* (24 January 2018).
2. Dennis, B. Trump budget seeks 23 percent cut at EPA, eliminating dozens of programs. *The Washington Post* (12 February 2018).
3. Hino, M., Benami, E. & Brooks, N. *Nat. Sustain.* <https://doi.org/10.1038/s41893-018-0142-9> (2018).
4. Liaw, A. & Wiener, M. *Classification and Regression by randomForest* (R news, 2002); https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf