ML for Public Policy Project Proposal
Tim Hannifan, Alec MacMillen, Quinn Underriner
April 30, 2019

Predicting Hazardous Waste Violations with EPA Data

Background

A plethora of evidence indicates that improper disposal, storage, and transportation of hazardous waste has a negative impact on the environment and public well-being. Contamination can occur via air, surface water, ground water, soils, sediments, and biological uptake. Rivers, their downstream waters, watersheds, surrounding populations, and the biological diversity that those areas support are threatened by a variety of contamination routes: groundwater seepage, surface runoff, bioaccumulation of hazardous compunts, and landfill leachate[1].

The implications of contaminations on public health are potentially catastrophic. The effects of exposure include cancer, genetic mutations, physiological malfunctions, physical deformations, and birth defects.[2]

There are many stakeholders in the realm of of hazardous waste production and storage. Large industries interests include those involved in the production of gold, aluminium, fertilizer, paper and textiles; small-scale industries include producers of chemicals, dyes/pigments, pesticides, and leather tanning chemicals. To these companies, compliance with EPA regulations is a cost, thus providing a disincentive to comply with federal and state statutes. On the other side is public interest and the security of the environment, to whom non-compliance imposes costs through negative impacts on health outcomes and taxpayer-funded cleanup.

Compliance is currently enforced three primary methods:

- Inspections: state and federal inspectors must assess each hazardous waste facility at least every two years. Facilities are either deemed compliant, or receive a violation specifying their transgression. The company must rectify the area of concern, then complete a follow-up inspection. Random inspections without advance notice or a search warrant are also permitted under current law.[3]
- Self-auditing: companies are exempt from gravity-based penalties if they self-report violations. This policy is intended to incentivise compliance by reducing costs to firms[4]
- Targeted inspections: see compliance monitoring strategy Defunct RCRA Watch List

[1] C.C. Barton, S.C. Schmitz, in Information Resources in Toxicology (Fourth Edition), 2009
[2] EPA Health and Ecological Hazards Caused by Hazardous Substances https://www.epa.gov/emergency-response/health-and-ecological-hazards-caused-hazardous-substances
[3] James A. Holtkamp, Linda W. Magleby, The Scope of EPA's Inspection Authority, Natural Resources & Environment, Vol. 5, No. 2, Administrative Law and Practice (Fall 1990), pp. 16-19, 47-48
[4] See Goal 3 in EPA's 3-Year Strategic Plan https://www.epa.gov/sites/production/files/2018-02/documents/fy-2018-2022-epa-strategic-plan.pdf

Project Goal

The aim of our analysis to create a targeted list of facilities that are likely to be found in violation of statutes within a two-year time span. The intention of creating this list is to increase the effectiveness of the EPA's inspection process by providing a ranked priority list that will identify likely violators, particularly those that are deemed 'high-value' targets based on the size of the company and potential cost to public health from their non-compliance.

The information produced by our analysis will be immediately applicable to federal and state inspectors of hazardous waste sites. Inspectors, who face time and resource constraints, would receive a list that provides the highest return on their time, as measured by the rate of violations per inspection. The public's interest would be served both by the decreased risk of health hazards, as well as the effectiveness of their tax dollars, part of which funds the EPA.

Data

The data we will use comes from the EPA's Enforcement and Compliance History Online (ECHO) system, which contains information from the Resource Conservation and Recovery Act Information System (RCRAInfo). This data is available from the EPA's website, at https://echo.epa.gov/tools/data-downloads/rcrainfo-download-summary. RCRAInfo tracks hazardous waste handlers, their activities, inspections of their activities, and enforcement and compliance actions taken. The data comes in six primary tables:

RCRA_FACILITIES – basic information about each RCRA site/hazardous waste handler, including name, location, and dummy variables that indicate what types of activities the facility engages in, its federal regulatory status, whether it transports hazardous waste, and whether it is an active site

RCRA_ENFORCEMENTS – count and description of enforcement actions taken against a given facility, the enforcement agency and date action taken, and dollar amounts of any settlements or punishments required

RCRA_EVALUATIONS – count and description of evaluations performed at a given facility, the evaluating agency and start date, and whether the evaluation resulted in a violation being found

RCRA_VIOLATIONS – detailed violation information, including the facility at which the violation occurred, its type, investigating agency, determination date, and the scheduled/actual date at which the facility brought the violation into compliance

RCRA_NAICS – the industrial activity classification/description of the facility, as defined by the Census Bureau (see below)

RCRA_VIOLATION/SNC HISTORY – dummy variables indicating whether the facility has unresolved violations and whether it has been designated as an unresolved significant noncomplier.

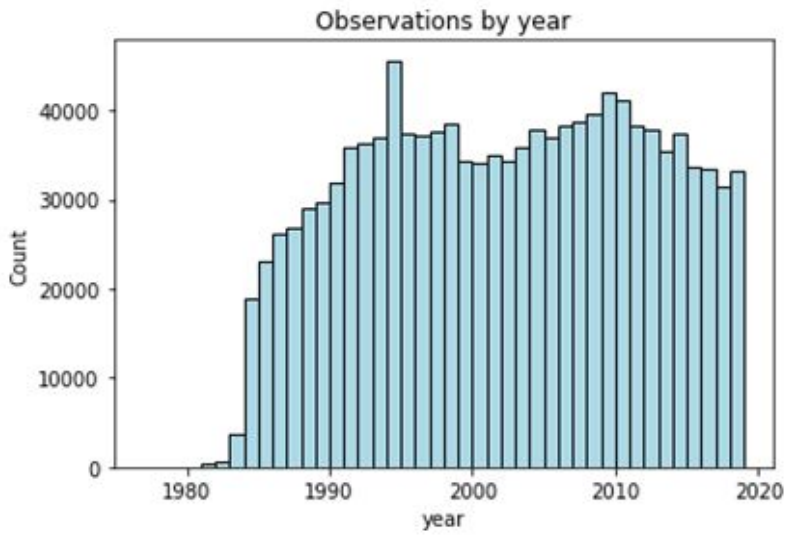
All tables contain a unique *ID_NUMBER* field that can be used to link all datasets. This ID indexes observations at the facility/site level, meaning that it uniquely identifies observations in the RCRA_FACILITIES dataset but may appear multiple times in other datasets (for example, *ID_NUMBER* may appear multiple times in RCRA_VIOLATIONS if the same facility commits multiple violations).

The main ancillary data source we will need is the Census Bureau's North American Industry Classification System (NAICS), which is a standard industrial coding system that defines a facility's activity type. This data is available at https://www.census.gov/eos/www/naics/. We will use this data in conjunction with the *NAICS_CODE* field in the RCRA_NAICS dataset to extract a description of each facility's activity type, which may be a useful predictor in the models we will build.

In thinking about potential features, there are some fields (like count of previous violations, indicators of unresolved violations and noncomplier status) that could potentially be strong indicators of whether an evaluation would result in a violation. It could also be informative to investigate the facility's status as a waste generator, hazardous waste transporter, and class of industrial activity.

A cursory review of the dataset provides an encouraging result: there appears to be a relatively consistent amount of data on a yearly basis over the span of the dataset. A simple merge of the RCRA_EVALUATIONS, RCRA_FACILITIES, and RCRA_NAICS tables results in a dataset of 1,225,500 observations, which are reasonably evenly distributed over the years 1984-2018 (with only a few outliers that fall outside this range). Within this dataset, roughly ⅔ of the observations (approximately 800,000) correspond to an evaluation that did not result in a found violation, while ⅓ did (roughly 400,000). Only 12,000 evaluations have a value "U" (undetermined).

Furthermore, there are very few null values of key variables. The only consistently null fields are *PMP_AMOUNT, FMP_AMOUNT, FSC_AMOUNT,* and *SCR_AMOUNT*, all found in the ENFORCEMENTS table. These fields represent the dollar amount of various fines and punishments levied on violating facilities, and it makes sense that there could be many null values as there are many possible enforcement actions beyond financial penalties.



Methodology

First, we will build a classification model that will predict whether or not a site with have a serious violation in the next two years. This will help with feature selection and help us see the potential and limitations of this data and different models. For this we will use a decision tree, logistic regression, random forests, and support vector machine models.

We will then tackle the temporal/ranking analysis. We are trying to prioritize which sites inspectors should visit first, given the predicted probability of a significant violation within the next two years. While the outputs from logistic regression may not be interpretable as literal probabilities of an event occurring, the magnitude of logistic estimations will provide a ranked order of most likely violators.

Validation

We plan on doing temporal holdouts for two-year periods given that two years is the maximum allowable time between compliance inspections. We want to minimize the False Discovery Rate (prediction error), because each false positive that is visited by the EPA represents time and resources that could have been used to inspect other facilities. Thus, the cost of false positives in this situation must be considered and we will seek to have as few false positives as possible.

Some caveats should be noted. We don't know the exact time specific nature of violations. Its possible that our model which reordered which facilities get inspected could do worse than we assume because the time of year in which the inspection occured matters. This model would need to be validated in the field by EPA officials to see if our assumption than a violation found late in the year would have existed earlier in the year holds. If our analysis is incorrect this would result in false positives.

The dates provided within the data may cause some issues with the accuracy of our time-bucketing. Within the data it's noted that the date listed might not be the date of the inspection that found the impropriety, but rather when test results were returned or when a legal determination was made. This important piece of data is subject to variance related to exogenous factors such as vagaries of the legal system and how backed up a governmental lab might be. Finally, we don't know the true number and location of non-compliers, only those who the EPA found to be non-compliers.

Policy Impact

The policy implications of our project are immediately apparent: federal and state agencies could mandate the use of a more intelligent system to improve the effectiveness of the inspection process. By taking the randomness out of selecting which facilities are targeted, we hope to improve the rate of return on inspector's time and taxpayer dollars.

The impact of implementation could be measured by the change in the rate of inspections resulting in a violation or citation. This analysis would require at least two years of data from before and after the implementation of any system or policy change. More immediate analysis could be performed, however, on incoming citation data within the two years after implementation. This early feedback would potentially prove helpful in calibrating the model and provide an early indication of its efficacy.

Appendix

```
Count of FOUND_VIOLATION (evaluation results)
---------------------------------------------

N      809186
Y      404336
U       11978


enforcements
 There are 332540 observations of 11 variables:

                    colname                     type  pct_null
0                 ID_NUMBER            <class 'str'>  0.000000
1           ACTIVITY_LOCATION            <class 'str'>  0.000000
2    ENFORCEMENT_IDENTIFIER            <class 'str'>  0.000000
3          ENFORCEMENT_TYPE            <class 'str'>  0.000000
4          ENFORCEMENT_DESC            <class 'str'>  0.000069
5     ENFORCEMENT_AGENCY                <class 'str'>  0.000000
6   ENFORCEMENT_ACTION_DATE            <class 'str'>  0.000000
7                PMP_AMOUNT  <class 'numpy.float64'>  0.955380
8                FMP_AMOUNT  <class 'numpy.float64'>  0.919856
9                FSC_AMOUNT  <class 'numpy.float64'>  0.997312
10               SCR_AMOUNT  <class 'numpy.float64'>  0.996364

Summary of numeric variables
      colname           mean   median  min  max      std_dev   count
0  PMP_AMOUNT   71880.810855  10000.0  NaN  NaN  768122.800148  332540
1  FMP_AMOUNT   49024.297915   7500.0  NaN  NaN  853275.571100  332540
2  FSC_AMOUNT  101109.412472  12198.5  NaN  NaN  754339.029781  332540
3  SCR_AMOUNT   34949.278031   9000.0  NaN  NaN  113164.781465  332540
-----------------------------------------------------------------------
evaluations
 There are 991196 observations of 8 variables:

                 colname          type  pct_null
0              ID_NUMBER  <class 'str'>       0.0
1      ACTIVITY_LOCATION  <class 'str'>       0.0
2  EVALUATION_IDENTIFIER  <class 'str'>       0.0
3        EVALUATION_TYPE  <class 'str'>       0.0
4        EVALUATION_DESC  <class 'str'>       0.0
5      EVALUATION_AGENCY  <class 'str'>       0.0
```

```
6  EVALUATION_START_DATE  <class 'str'>       0.0
7         FOUND_VIOLATION  <class 'str'>       0.0

Summary of numeric variables
 None
------------------------------------------------------------------------
facilities
 There are 1041254 observations of 15 variables:

                    colname                    type       pct_null
0                 ID_NUMBER             <class 'str'>  0.000000e+00
1             FACILITY_NAME             <class 'str'>  5.762283e-06
2         ACTIVITY_LOCATION            <class 'str'>  0.000000e+00
3          FULL_ENFORCEMENT            <class 'str'>  0.000000e+00
4   HREPORT_UNIVERSE_RECORD            <class 'str'>  9.315691e-05
5            STREET_ADDRESS            <class 'str'>  1.306117e-04
6                 CITY_NAME             <class 'str'>  9.795881e-05
7               STATE_CODE             <class 'str'>  9.603805e-07
8                 ZIP_CODE             <class 'str'>  1.930365e-04
9               LATITUDE83  <class 'numpy.float64'>  5.653500e-01
10             LONGITUDE83  <class 'numpy.float64'>  5.653500e-01
11      FED_WASTE_GENERATOR            <class 'str'>  8.250629e-03
12              TRANSPORTER             <class 'str'>  2.823519e-04
13              ACTIVE_SITE             <class 'str'>  0.000000e+00
14          OPERATING_TSDF         <class 'str'>  0.000000e+00

Summary of numeric variables
 None
------------------------------------------------------------------------
naics
 There are 426042 observations of 3 variables:

           colname          type  pct_null
0          ID_NUMBER  <class 'str'>     0.0
1  ACTIVITY_LOCATION  <class 'str'>     0.0
2          NAICS_CODE  <class 'str'>    0.0

Summary of numeric variables
 None
------------------------------------------------------------------------
violations
 There are 594380 observations of 8 variables:

                colname          type  pct_null
0                 ID_NUMBER  <class 'str'>  0.000000
```

```
1          ACTIVITY_LOCATION  <class 'str'>  0.000000
2             VIOLATION_TYPE  <class 'str'>  0.000000
3        VIOLATION_TYPE_DESC  <class 'str'>  0.000000
4  VIOL_DETERMINED_BY_AGENCY  <class 'str'>  0.000000
5  DATE_VIOLATION_DETERMINED  <class 'str'>  0.000000
6            ACTUAL_RTC_DATE  <class 'str'>  0.021666
7  SCHEDULED_COMPLIANCE_DATE  <class 'str'>  0.000000

Summary of numeric variables
 None
------------------------------------------------------------------------
viosnc_history
 There are 2318155 observations of 5 variables:

            colname                type  pct_null
0           ID_NUMBER        <class 'str'>    0.0
1  ACTIVITY_LOCATION        <class 'str'>    0.0
2           YRMONTH  <class 'numpy.int64'>    0.0
3           VIO_FLAG         <class 'str'>    0.0
4           SNC_FLAG         <class 'str'>    0.0

Summary of numeric variables
 None
------------------------------------------------------------------------
```