EPA1 Project Proposal

Carly Jones (carlyjon), Wenyu Hunag (wenyuh), Mike Chen (xiangtic)

1 Background and Goals

1.1 Problem Statement and Goal

The New York State Department of Environmental Conservation (NYSDEC) is the entity authorized by the United States Environmental Protection Agency (EPA) to enforce federal regulations for the handling of hazardous waste in New York. With more than 118,527 registered hazardous waste facilities in the state, NYSDEC must use the information at their disposal to determine which facilities will be chosen for inspection each year. The inspections are time consuming and costly for both the department and the industry partner which manages the facility.

The primary goal of our project is to develop a system for predicting which of the many hazardous waste facilities in New York are likely to incur one or more serious violations within the next year. Other goals include ensuring that each facility is inspected at least once every 10 years, maintaining an equitable distribution of inspections so that individual industry partners are not unduly burdened by constant inspections, and encouraging prioritization of inspections for facilities in densely populated or otherwise high-risk communities.

The output of our project will inform NYSDEC's programmatic decision-making and, most importantly, decrease the overall number of serious violations through enforcement and informed mitigation.

1.2 Background

NYSDEC is responsible for implementing the Resource Conservation and Recovery Act (RCRA) program for New York State. NYSDEC's major responsibility is to enforce federal and state regulations regarding the management of hazardous wastes in New York as well as conducting regular inspections.

The inspection program, which is managed through nine Regional Offices, performs approximately 700 inspections statewide each year. Among them, approximately 55% of the facilities are found to be substantially in compliance with state regulations. Approximately 40% of facilities are Secondary Violators (SVs), and the remaining 5% are Significant Non-Compilers (SNCs) [3]. See Appendix A for a detailed description of the inspection process.

The scope of inspection includes:

- **Generators**: facilities that produce hazardous wastes. Depending on the volume of hazardous waste each generator produces, generators are further grouped into three categories (a) Very Small Quantity Generators (VSQGs); (b) Small Quantity Generators (SQGs); (c) Large Quantity Generators (LQGs) [1].
- **Treatment, storage or disposal facilities (TSDF)**: sites which use land and structures for treating, storing, and disposing of hazardous wastes.
- **Transporters**: people engaged in the off-site transportation of hazardous waste within the US, which includes shipment from a Generator or property to a TSDF [2].

1.3 Baseline and Existing Solution

Compliance monitoring system is the system set by the Office of Enforcement and Compliance Assurance (OECA) to monitor whether certain facilities have violations against applicable law. The goal of the compliance monitoring is to ensure hazardous waste are managed properly so it will not endanger human health and the environment, given the limited resources. The overall frequency of coverage is inspecting 20% LQGs and over 50% TSDFs every year so that all LQGs will get inspected and monitored in 5 years and TSDFs in 2 years. The system requires the coordination between the State and the Region. Both State and Region have the rights to investigate and inspect facilities, but the State has the primary responsibility and regions should ensure and offer help to cover all facilities.

Besides fixed frequency of coverage, both State and Regopm follow rules of priority to decide the order or the urgency to inspect certain facilities. Here are the rules in order of priority [4]:

1. National Priority Sectors
2. Never-inspected LQGs
3. Non-notifier facilities believed to generate hazardous waste in quantities that would require notification.
4. Persons that generate, transport, treat, store, or dispose of significant quantities of hazardous waste, particularly those in proximity to population centers, areas with environmental justice concerns, or environmentally sensitive areas
5. Repeat violators
6. Facilities with complex operations or processes that increase the likelihood of missing waste streams or making improper exemption determinations
7. Facilities that are the subject of citizen complaints.

For example, OECA is less likely to inspect a facility that has a track-record of compliance and whose monitoring activities meet the minimum standard qualitatively. However, one caveat is that OECA only specifies what they prioritize but did not give specific instructions. It relies more on inspectors' subjective decisions. This again displays the necessity of a model that gives recommendations of inspections based on the factors they prioritize.

2 Data Description and EDA

2.1 Datasets

There are in total 7 schemas given, of which 2 schemas are specific to New York State. A short description for each is presented below (see Appendix B for additional information):

- RCRA contains inspection, violation, and enforcement data related to RCRA, as well as information about facilities and handlers of hazardous waste. This is the main schema that we're looking at because table `cmecomp3` offers data related to past violations, which can be a good reference for our model.
- FEC contains federal enforcement and compliance (FE&C) data from the Integrated Compliance Information System (ICIS).
- FRS allows for linking facilities between data provided by ICIS and RCRA.

- AIR contains emissions, compliance, and enforcement data on stationary sources of air pollution.
- NPDES contians data on the compliance and enforcement status of facilities regulated by the National Pollutant Discharge Elimination System (NPDES).
- NYSDEC-REPORTS includes information from reports filed annually by LQGS as well as TSDF in the state of New York.
- MANIFEST contains information about hazardous waste shipments to, from, or within the state of New York.

2.2 EDA Results

For exploratory data analysis, we mainly focused on inspection-related data from 2014 to 2016 in the state of New York extracted from the full RCRA dataset (specifically, the `temp_tables.cmecomp3 _ny_2014_16` in `epa1_database`), which contains information on 10,257 inspections conducted over this three year period.

Our initial review of inspection records from 2014 to 2016 provides a few key insights. First, as indicated by the left of Fig. 1, the majority of facility inspections result in some kind of violation (82.49%). This leads us to believe that simply predicting a violation (of any kind) will not be enough to help NYSDEC effectively prioritize their inspection roster.

Second, certain kinds of violations are more prevalent than others (see Table 1). Additional qualitative research on the potential consequences of the 42 types of violations represented in our inspection records will allow us to categorize them by their severity, as measured by their potential economic, community, or environmental impact.

Third, the range of the number of violations made by individual facilities over the three year periods large (between 0 and 192, see right of Fig.1 and Table 2), telling us that certain facilities are worse violators than others. Our analysis will need to take this into account, and adequately balance targeted enforcement with adequate coverage of all facilities over a reasonable time frame, perhaps 10 years.

Enforcement is the required compliance of violated laws. By observing the enforcement distribution of facilities from 2014 to 2016 in Table 3, we can see that over 40% of violations have written informal enforcement, which means that an agency representative notifies the waste sites that they violated certain terms and advises them to make corrections and by when to correct them. This is not a formal action. In other words, this is not a significant non-complier. Logically, given the limited resources, the model of recommendation should advise agency representatives to inspect facilities with formal actions, because those facilities have higher risk to hazard humans health and environment. The enforcement type allows us to classify facilities into significant non-compliers and non-significant non-compliers, which will be helpful to give recommendation of inspections.

3 Proposed Analysis and Validation Methods

Given our stated goals, we propose initially formulating our analysis as a binary classification problem. Using appropriate methods, we will develop a model to predict which facilities in New York State are most likely to incur one or more severe violations within the next year.

We will consider and test several binary labels, all of which are either already available in our dataset or will be available with feature engineering:

- Met minimum threshold or percentile for total number of violations within the past 1 year/3 years/5 years
- Met minimum threshold or percentile for total number of severe violations (filtered by violation type) within the past 1 year/3 years/5 years
- Designated as a significant non-complier within the past 1 year/3 years/5 years
- Met minimum threshold or percentile for penalties incurred for violations within the past 1 year/3 years/5 years
- Met minimum threshold or percentile for an engineered violation impact score for violations within the past 1 year/3 years/5 years

Depending on the way we "window" the time-scaled labels proposed above, we may have significantly fewer positive observations (severe violators) than negative observations (non-severe violators). For this reason, we may need to upsample the positive observations prior to training.

We plan to evaluate several classification algorithms, including logistic regression, boosted decision trees, and random forest. Additional exploratory data analysis will help us to better understand our feature set, and its potential separability, before selecting any additional methods to test.

We will validate our analysis with "windowed" crossfold training and test sets which use a specified number (perhaps 1, 3, or 5) of previous years' data to predict the following year. The following year, in this case, becomes the test set. Calculations of precision and recall will help us understand the test accuracy of the model, and a receiver operating characteristic (ROC) curve will help us visualize the model's discriminatory power. In addition, we can calculation the area under the curve (AUC) as a metric of performance.

Another possible problem formulation that we may explore is a regression problem. As shown by the EDA, more than 80% of the facilities committed some kind of violation. Thus, for NYSDEC, the number of predicted violations within the next year might be a more useful metric in guiding inspection efforts. Algorithms that we use include boosted regression trees, and the evaluation metric would be mean squared error.

Besides simply measuring the accuracy of our model, we would like to use additional metrics to assess how well our model output supports our goals of comprehensiveness (inspecting all facilities within a certain time frame), fairness (distributing inspections between facilities), and community health (explicitly prioritizing facilities in densely populated or otherwise high-risk areas). Overall, we hope to use these metrics to choose a model that best balances these competing objectives.

4 Caveats

The provided datasets have the following limitations:

1. The training and testing data distribution might be different. In other words, the model will be trained on data related to facilities that have been inspected in the past, but the

prediction is made the entire population of facilities.

2. It's difficult to quantify the severity of violations as we cannot estimate the true impact of an incident that might have happened if inspections were not conducted.
3. As discussed in Section 1.3, the underlying mechanism through which OECA selects the facilities to examine can be very subjective. Moreover, it's uncertain whether the mechanism is consistent across time. Since the training data is facilities that were inspected in the past, we might have a distribution shift overtime (i.e., the underlying distribution of facilities that are inspected changes as the selection mechanism changes), which make model-training more challenging.

5 Ethical Considerations

As stated above, hazardous waste incidents have the potential to damage the economy, community, and environment in which they occur. As we develop our system, we will continuously evaluate our progress on the goals of limiting the potential impact of a hazardous waste incident on individual people, protecting important natural resources, and distributing the financial burden of inspections. We will use data from the American Community Survey (ACS) to support these metrics.

Transparency around the way that our system supports inspection prioritization will also be important, given that inspections are time-intensive, costly, and publicly-funded. Communicating with industry and the general public about the methods of evaluation our system uses will also support partnering and buy-in.

6 Policy Recommendations

Through this project, we hope to help with the inspection decisions of agency representatives so that they can inspect the facilities with high risks of violations. Given the limited time and resources, we hope to maximize the efficiency of inspection coverage in support of our stated goals. In this way, NYSDEC and EPA can be guided to boost their efficiency; industries or facilities can be properly monitored and hence decrease the amount of violations and their impact on the economy, community, and environment.

Our analysis may also reveal certain industrial methods or equipment that consistently contribute to high risk of a violation. This information can inform EPA and NYSDEC policy implementations, training programs, and industry outreach.

To evaluate whether what we have proposed has the desired impact, we hope to send out surveys to agency representatives about their experience using the model. For example, the differences between recommendation given by the model and their subjective decisions. Post-implementation, we may need to conduct a A/B Test-like field-experiment where half of the facilities are selected by NYSDEC and the other half are supplied by our model and compare the results. We can check after applying the model, whether the amount of improperly generated or handled hazardous waste decrease and whether facilities improve process and operations by substituting non-hazardous waste or other means.

Appendix

A Inspection Process

The inspection program conducted by NYSDEC has the following steps:

1. Hazardous waste handling sites and areas are physically inspected (e.g., tanks, container storage areas, incinerators, etc.). Reports written by the facilities are also reviewed.
2. After the physical inspection and record/report reviews, the NYSDEC inspector would complete an inspection report and makes a compliance determination.
3. If a violation is confirmed, depending on its severity, a facility would either be classified as a Significant Non-Complier (SNC) or a Secondary Violator (SV). SNCs may be resolved as civil or criminal enforcement actions, while SVs are resolved using a Notice of Violation.

B Additional Dataset Information

B.1 RCRA

This dataset is published online by EPA regarding the facilities, past inspections, violations and enforcement. The primary key of this schema is `rcra_id` which is a unique identifier assigned to all facilities include generators, TSDFs, transporters etc.

B.2 FEC

This set of data records the federal administrative and federal judicial EPA enforcement cases. Therefore, the primary key is `case_number`, `activity_id` (both keys are used in this schema only) and in some tables, `registry_id` (which is the unique `facility_id` used in all ICIS tables). It documents general case information (case no., case name), violation information, defendant information, milestone dates, and penalty amount. This schema might be useful to determine the relationship between case information (the pollutants involved, the severity etc.) to the penalty amount. When joined with other tables using `facility_id`, we can access more detailed case information of waste sites.

B.3 Facility Registry Service (FRS)

The Enforcement and Compliance History Online (ECHO) system published FRS as the linking table between ICIS and RCRA tables. The most useful function of this table is to map between RCRA `handler_id` and `registry_id` / `facility_id` used in ICIS. When one of the primary key program system acronym is RCRAInfo, program `system_id` is just the `handler_id`. This is an important table we can use to integrate information of RCRA and ICIS tables.

B.4 Air

This dataset is published by EPA. It contains emission, compliance, and enforcement data on stationary sources of air pollution. The unique feature of this set of tables is that it focuses on air pollution instead of hazardous waste so it can be used as complimentary material for air pollution violations. It has general source information, past violations, air test history etc.

The primary key is PGM_SYS_ID (program system identifier that uniquely identifies each air source).

B.5 NPDES

This table tracks the compliance and enforcement status of facilities under Clean Water Act. In other words, this set of tables focuses on water pollution. There are in total 3 kinds of violations: schedule violations, effluent violations, and DMR non-receipt violations. For each of them, there are one or two automatic inspection machines. Since water is critical to people's health conditions, this table could be used to relate with health data and be a great factor in model. The primary key is npdes_id (internal id) and facility_id.

B.6 NYSDEC-Report

This dataset contains information of LQGS and waste generated in LQGs annually and waste received by TSDFs in New York state. The primary key is handler_id which is also used in RCRAInfo. We are able to identify and observe the distribution of waste in facilities in New York.

B.7 Manifest

This dataset tracks the hazardous waste shipments to, from or within New York state. It is case-wise which means whenever a new shipment gets started, it will be added to the database. It has information like generators, transporters, TSDFs, waste amount, waste types involved in transportation. All facilities involved are identified by handler_id from RCRAInfo.

C EDA Plots and Figures

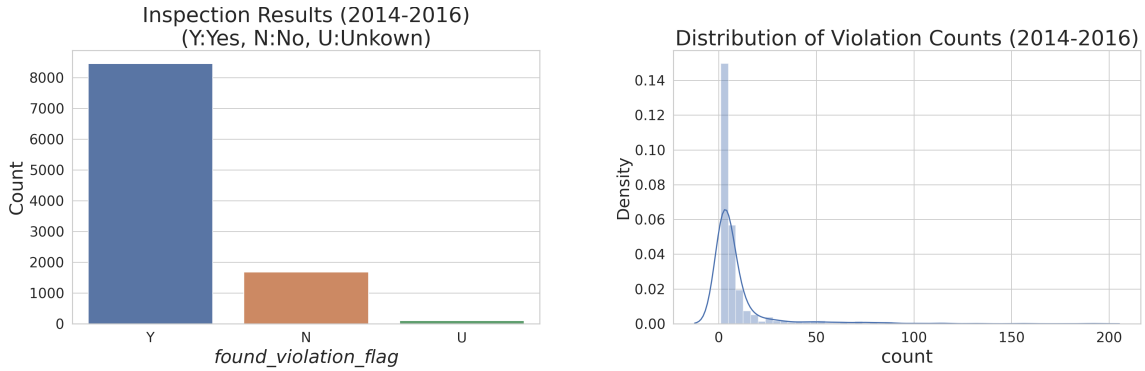

Figure 1: Left: inspection result breakdown. Right: distribution of the number of violations

	Violation Code	Violation Short Description	Count
1	262.C	Generators - Pre-transport	1579
2	273.B	Universal Waste - Small Quantity Handlers	1329
3	XXS	State Statute or Regulation	1292
4	265.C	TSD IS-Preparedness and Prevention	564
5	265.I	TSD IS-Container Use and Management	493
6	265.B	TSD IS-General Facility Standards	473
7	265.D	TSD IS-Contingency Plan and Emergency Procedures	434
8	261.A	Listing - General	385
9	262.A	Generators - General	271
10	279.C	Used Oil - Generators	220

Table 1: Top types of violation found in New York from 2014 to 2016

	Handler ID	Violation Count
1	NYD981140387	192
2	NYD002080034	156
3	NYD001212752	132
4	NYR000143396	117
5	NYD000379248	114
6	NYD057381535	111
7	NYR000213280	102
8	NYR000150672	99
9	NYR000220525	90
10	NYR000218677	88

Table 2: 10 Handlers who committed the most violations from 2014 to 2016 in New York

Enforcement Code	Enforcement Description	Count	Percentage	Formal
120	Written Informal	3504	41.41	**N**
140	Letter of Intent to Initiate Enforcement Action	2120	25.06	N
210	Initial 3008(a) Compliance	1537	18.17	Y
310	Final 3008(a) Compliance Order	1121	13.25	Y
None	NA	103	1.22	NA
110	Verbal Informal	75	0.89	N
250	Field Citation	1	0.01	N

Table 3: Enforcement type distribution of facilities with violations from 2014 to 2016 in New York

References

[1] EPA. Categories of Hazardous Waste Generators. (Online; accessed 24-September-2020). URL: https://www.epa.gov/hwgenerators/categories-hazardous-waste-generators.

[2] EPA. Hazardous Waste Transportation. (Online; accessed 24-September-2020). URL: https://www.epa.gov/hw/hazardous-waste-transportation.

[3] NYSDEC. Compliance Inspection Program for Hazardous Waste. (Online; accessed 24-September-2020). URL: https://www.dec.ny.gov/chemical/8773.html.

[4] EPA OECA. Compliance Monitoring Strategy for the Resource Conservation and Recovery Act (RCRA) Subtitle C Program. (Online; accessed 24-September-2020). URL: https://www.epa.gov/compliance/resource-conservation-and-recovery-act-rcra-compliance-monitoring.