The Distributive Effects of Risk Prediction in Environmental Compliance: Algorithmic Design, Environmental Justice, and Public Policy

Elinor Benami
Virginia Tech
elinor@vt.edu

Reid Whitaker
University of California Berkeley
reidw@stanford.edu

Vincent La
Stanford University
vincela@law.stanford.edu

Hongjin Lin
Stanford University
hongjinl@law.stanford.edu

Brandon R. Anderson
Stanford University
banderson@law.stanford.edu

Daniel E. Ho
Stanford University
dho@law.stanford.edu

ABSTRACT

Government agencies are embracing machine learning to support a variety of resource allocation decisions. The U.S. Environmental Protection Agency (EPA), for example, has engaged academic research labs to test the use of machine learning in support of an important national initiative to reduce Clean Water Act violations. We evaluate prototypical risk prediction models that can support compliance interventions and demonstrate how critical algorithmic design choices can generate or mitigate disparate impact in environmental enforcement. First, we show that the definition of which facilities to focus on through this national compliance initiative hinges on arbitrary differences in state-level permitting schemes, causing a shift in environmental protection away from areas with more minority populations. Second, the policy objective to reduce the noncompliance *rate* is encoded in a classification model, which does not account for the extent of pollution beyond the permitted limit. We hence compare allocation schemes between regression and classification, and show that the latter directs attention towards facilities in more rural and white areas. Overall, our study illustrates that as machine learning enters government, algorithmic design can both embed and elucidate sources of administrative policy discretion with discernable distributional consequences.

CCS CONCEPTS

• **Social and professional topics** → **Governmental regulations**; • **Applied computing** → **Law**; **Computing in government**; • **Computing methodologies** → *Machine learning*; • **Human-centered computing** → *Interaction design.*

KEYWORDS

risk models, government, environmental protection, fairness, environmental justice

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445873

ACM Reference Format:
Elinor Benami, Reid Whitaker, Vincent La, Hongjin Lin, Brandon R. Anderson, and Daniel E. Ho. 2021. The Distributive Effects of Risk Prediction in Environmental Compliance: Algorithmic Design, Environmental Justice, and Public Policy. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3442188.3445873

1 INTRODUCTION

Governments are rapidly experimenting with machine learning for public policy, raising significant questions about accountability, fairness, and governance [12, 29, 38]. One emerging application area is in environmental sustainability [36]. Serious noncompliance exists across environmental programs [19], and machine learning offers the promise to help predict sources of noncompliance and thereby target environmental compliance efforts [20, 22].

This paper considers the case of a National Compliance Initiative (NCI) for the Environmental Protection Agency (EPA). This NCI aims to reduce significant noncompliance (SNC) under the Clean Water Act, the nation's premier piece of legislation to protect the waterways of the United States. EPA's goal is to reduce the SNC rate by 50% from 2019 to 2022, relative to a baseline from 2018 [35]. The NCI was pioneering in scope relative to previous enforcement efforts that tended to be sector or facility-type specific. Most importantly, as we document below, EPA extended the scope of enforcement priorities by including both major and minor facilities and engaged in a comprehensive assessment of programmatic efforts to secure compliance [3]. In support of this NCI and broader environmental compliance goals, the EPA has laudably engaged the academic community to discern how machine learning methods can contribute.

These engagements offer a fruitful opportunity to study distributive implications in this important public policy domain. We show that two key elements of policy and algorithmic design may have considerable distributive effects, influencing who bears the burden of excess pollution and how intensely. First, the NCI was designed to measure compliance across a subset of permitted facilities under the Clean Water Act. Notwithstanding the major expansion of coverage under the NCI to include minor facilities, the same type of facility (e.g., a wastewater treatment plant) may be included or excluded from the purview of the NCI due to variations in how states implement their pollution permitting process. We show that this decision in effect converts an ambitious and well-intentioned *national* compliance initiative into a more of a *patchwork* compliance initiative that functionally concentrates on only a handful of states and communities. In effect, environmental federalism impedes the national goal. Second, the NCI targets the *rate* of noncompliance, regardless of how much individual facilities discharge above their permitted limits. In machine learning terms, this policy choice leads to *classification* of facilities into whether they are likely to be in SNC status, not *regression* of predicted discharges above the limit. We show that this objective shifts resources away from the most severe violators in higher minority areas towards smaller facilities in areas with fewer minorities as a share of the overall population. We also compare these effects against an "oracle test" and show that ML models may increase or decrease disparate impact relative to the case with full knowledge of realized outcomes.

Our study illustrates that fair algorithmic design will increasingly be intertwined with policy discretion. While government use of predictive algorithms can unwittingly reinforce prior discretionary policy choices, formalization in algorithmic design can also provide a chance to study the potential for disparate impact associated with such policy choices.

The rest of the paper proceeds as follows. Section 2 discusses the related literature, and Section 3 provides background on the policy setting. Section 4 details our methods, particularly the development of the risk model and tests for demographic bias. Section 5 provides results and Section 6 concludes.

2 RELATED LITERATURE

Our work contributes to several distinct literatures. First, as the public sector has rapidly adopted machine learning systems [12, 29], core questions have focused on the accountability of algorithmic decision tools in the face of public law constraints [10, 24, 26]. While much attention has focused on the use of risk assessment scores and facial recognition technology in criminal justice, far fewer in-depth investigations and case studies exist of the adoption of such decision tools in civil justice [for important exceptions, see 6, 7, 17]. Chouldechova et al., for instance, examines fairness of an algorithmic decision making tool for child welfare determinations. Few works in the the FATML community concern the use of algorithms in environmental sustainability, however, and our study provides an in-depth case study of algorithmic design, policy, and accountability in this policy domain. Our results also contribute to questions about the perceived trade-off between accuracy and explainability in machine learning. Despite the fact that our data draws on rich information from hundreds of millions of EPA records, we show that relatively simple models provide much of the predictive performance in this particular domain.

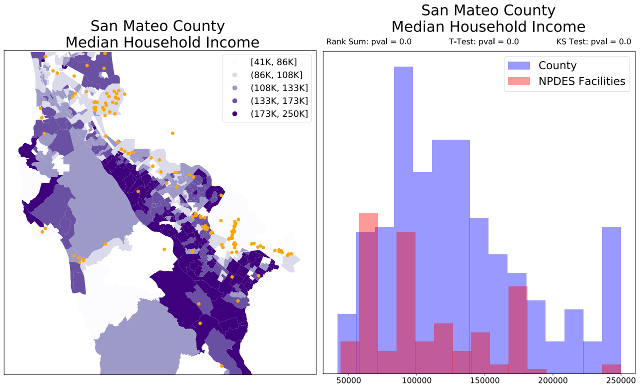


Figure 1: Census Block Groups with NPDES permitted facilities tend to fall in the lower tail of the income distribution in a given county, as illustrated by mapping the location of NPDES permittees (in map, orange dots; in histogram, salmon-colored bins) within the distribution of median incomes, in the example case of San Mateo County, California.

Second, our study relates to a growing literature, albeit one that is largely disconnected from machine learning, about the distributional consequences of environmental resource allocation. Wikstrom et al. employ the CalEnviroScreen EJ monitoring tool – a tool that combines data on environmental burdens with sociodemographic data – to assess how water resource allocation policies in the form of cutbacks during drought differentially affect minority communities highly ranked in the EJ assessment. Maguire and Sheriff characterize the evolution of the term of Environmental Justice, as well as techniques to study and surface the possibility disproportionate environmental harms that may arise from the regulatory rulemaking. The economists Banzhaf et al. review the spatial nature of environmental justice concerns and characterize the multiple mechanisms that can give rise to disproportionate harm and exposure landing on some communities over others. Much of this literature has attempted to understand the causal factors of environmental injustice, often using retrospective observational studies. Yet, as the concept of environmental justice becomes institutionalized, it is also critical to understand what can be done about these disparities prospectively. EPA established the Office of Environmental Equity in 1994 to address concerns that "racial minority and low-income populations bear a higher environmental risk burden than the general population." In the same year, all federal agencies were tasked to examine the disproportionate harms their programs may have low-income and minority communities through an executive order on environmental justice [8]. Our case study shows how algorithmic decisions that are actively being developed may exacerbate or mitigate disparate impact depending on key design decisions.

Third, related to the broader environmental justice literature, an extensive body of work has specifically investigated disparities in the siting of polluting activities. Figure 1, for instance, speaks to some of these concerns by plotting the Clean Water Act permitted facilities in San Mateo County, California, illustrating that such facilities are disproportionately sited in lower income areas. Banzhaf et al. review the complex causal dynamics driving these disparities. We contribute to this literature by examining whether, conditional on siting decisions, the implementation of the NCI may have a further distributive effect on which communities are protected from pollution.

Last, our study contributes to the literature on regulatory enforcement. Hindin and Silberman focuses on mechanisms to improve rule design to promote compliance. Konisky examines whether the number of enforcement actions taken by state EPAs is correlated with demographics at the county-level. And many studies have speculated that algorithmic decision tools may improve the accuracy and consistency of regulatory enforcement. Hino et al., for instance, illustrate how protoyipical machine learning models applied on publicly available data about permitted facilities under the Clean Water Act could help more effectively target facilities at risk of violation. Engstrom and Ho consider whether algorithmic decision tools may improve the quality of enforcement actions by the Securities and Exchange Commission and determinations by the Social Security Administration, where a central concern of administrative law has been about the consistency of decision making.

This paper seeks to bridge these distinct literatures — spanning economics, legal studies, environmental studies, geography, public administration, and machine learning — to better understand how algorithmic design affects how and where resources to manage environmental quality are allocated.

3 POLICY BACKGROUND

The Clean Water Act (CWA) is the principal legislation governing surface water pollution in the United States [31]. Its primary objective is to "restore and maintain the chemical, physical, and biological integrity of the Nation's waters." The CWA endows the EPA with the authority to implement pollution control programs as well as set standards for wastewater and surface water quality. One of the key ways EPA implements these tasks is through the National Pollutant Discharge Elimination System (NPDES) permit program.

For any discharge from a point source (e.g., a pipe) into U.S. surface waters, facilities must apply for a NPDES permit. That permit specifies the conditions under which facilities can discharge, setting limits on water quality parameters (e.g., the quantity or concentration of nitrogen, phosphorous, or metals; acidity (pH); and temperature). Discharges in violation of the CWA are subject to civil and criminal penalties, with the goal of incentivizing facilities to adopt advanced water treatment and pollution reduction technologies. The primary monitoring mechanism for the CWA comes through reporting obligations on facilities. Permitted facilities are required to self-report information on the water quality of their discharges, typically monthly, to environmental authorities. These "discharge monitoring reports" (DMRs) provide information on the results of water quality tests on features such as temperature, pH values, and the quantity and concentration of, for example, solids that can transport pollutants or inhibit marine life. As with much U.S. environmental law, the CWA is an arrangement of cooperative federalism. For 47 states (and one territory), EPA has delegated the authority to state environmental agencies to administer the NPDES programs.

Despite the ambition of the CWA, noncompliance under the CWA remains pervasive. Based on self-reported DMRs, between 60-75% of facilities are in noncompliance each year, contributing to water quality impairment that can render streams and rivers unswimmable or unfishable. Because of these patterns, EPA's five-year strategic plan highlighted the CWA as an area of enforcement focus and established the goal of cutting significant noncompliance in half over a three year term. In support of its broader enforcement goals, EPA has already engaged multiple academic labs to develop machine learning methods that leverage the large-scale administrative data within the EPA – estimated by one analyst to be "the largest federal government database outside those of the Internal Revenue Service" – and facilitate early interventions (enforcement actions). Rather than reacting to facilities that are already noncompliant, the EPA aims to use predictive risk assessment techniques to target and then prevent facilities from becoming noncompliant in the first place. We now articulate two major policy and algorithmic design decisions in the deployment of machine learning for this initiative.

3.1 Targeted Population

For the purposes of the NCI, the CWA compliance rate is calculated based on the subset of CWA-regulated facilities that submit DMRs to federal EPA. Beginning in 2016, states were required to push electronic versions of the DMRs to the federal EPA [13]. This rule also expanded reporting obligations from primarily 'major' to include 'minor' facilities – two classes of facilities distinguished principally based on daily flow. Previously, only major permits typically reported DMRs and the reporting was not necessarily conducted electronically, but the inclusion of minor permits was a substantial and important expansion of the scope of enforcement priorities (by nearly seven-fold, as indicated in Table 1). At the same time, the first phase of this "electronic reporting rule" was made applicable only to "individual permits," which cover a specific discharging entity (e.g., a single wastewater treatment plant). States are currently not required to push DMRs for so-called "general permits," which are permits to cover multiple dischargers engaged in similar activities and with similar types of effluent [18], and will not be required to until the next phase of the electronic reporting rule's implementation, scheduled for 2022. However, whether a wastewater treatment facility is licensed under an individual or general permit is largely contingent on state-level choices: the same type of facility may be permitted as either general or individual depending on the historical permitting decision systems made by states. Moreover, many general permittees do in fact submit DMRs to EPA, making it possible to study what impact the decision to delay the inclusion of general permittees implies for the distribution of environmental protection.

3.2 Noncompliance Rate as Objective

The second major dimension of policy design rests on the ultimate policy objective, i.e., the focus on the reducing the *rate* of "significant noncompliance" (SNC). The SNC designation refers to the most serious class of CWA violations considered to pose a threat to U.S. waterways. As outlined in a 1995 EPA legal memorandum, several criteria may result in a facility falling into SNC [34], including failure to report discharges within 30 days of expected dates; or persistent, considerable excess of permitted limits.

As Table 1 indicates, approximately 20,000 facilities (~30%) were classified as in SNC in fiscal year 2018 (henceforth FY2018). Of the five key types of SNC (outlined in [16]), this paper focuses on effluent violations, which constitute 20% of the facilities in SNC status, and that, furthermore, pose a directly measurable threat to water quality. In broad brushstrokes, a facility is considered as having an effluent-related SNC if one of their permitted discharges exceeds its permitted limit by any amount four times within two consecutive quarters or exceeds a predetermined threshold twice within two quarters [34]. The Code of Federal Regulations [9] and the 1995 memo [34] note that the predetermined SNC threshold beyond the permitted limit for conventional pollutants (e.g., Nitrogen, Phosphorous, total suspended solids, detergents, oils, and total organic carbon) is 40% and for toxic pollutants (e.g., most metals, cyanide, and toxic organic compounds) is 20%. For simplicity of exposition, we will use SNC to refer to effluent SNC for the remainder of the paper unless explicitly indicated otherwise.

Status	General	Indiv. Major	Indiv. Minor
Total SNC	**11,155**	**1,147**	**7,366**
DMR Nonreceipt SNC	10,110	472	4,057
Effluent SNC	953	503	2,480
Other SNC	92	172	829
Non-SNC	**26,649**	**2,472**	**17,316**
Total	**37,804**	**3,619**	**24,682**
SNC Rate	29.5%	31.7%	29.8%

Table 1: Types and Quantities of Significant Noncompliance among General and Individual Permittees for FY2018

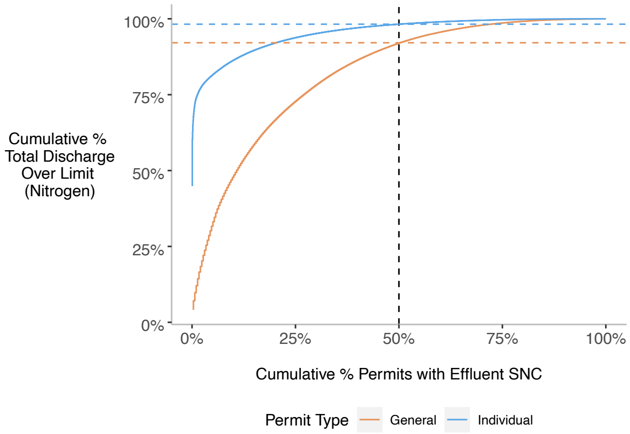


Figure 2: Some permits account for a much higher share of overall effluent exceedances than others. For example, where the dashed lines indicate half of all evaluated permits per permit type and permits are sorted by their percent exceedance (high to low), the at-right half of individual (general) permits closest to the effluent SNC limit account for only 1.8% (7.9%) of the aggregate percent over limit for the 1,590 (297) permits recording nitrogen discharges and flagged as in effluent SNC in FY2018.

We examine the implications of using a rate as the policy objective in greater detail below because the current definition of "significant noncompliance" does not distinguish between the extent of effluent violations once the initial thresholds are met. For example, Figure 2 plots the cumulative distribution of how much permittees reporting nitrogen effluent exceed their limits, in percent-above-limit terms, sorted by their exceedance percent. Each of the permits would contribute equally to the current NCI goal of reducing the SNC rate upon returning to compliance. However, some permits exceed their permits by a much larger margin than others: the lowest half of the individual (general) facilities flagged as with an effluent SNC account for 1.8% (7.9%) of the aggregate percent over limit for the 1,590 (297) permits recording nitrogen discharges and flagged as in effluent SNC in FY2018. As seen on the left side of the graph, a single individual permit contributes to 40% of the total across all permits. In other words, a classification lens considers all these permits equal, although one could reasonably think that permittees who far exceed their limits would generate a much higher negative impact than those marginally overstepping their limits.

4 METHODS

4.1 Data

This work draws upon three main sources of data. First, we extract information about historical discharge volumes, compliance history, and permit-level variables (e.g., individual versus general) from EPA's Integrated Compliance Information System (ICIS) on NPDES permits [32]. A key data source within ICIS rests in the over 270 million records from Discharge Monitoring Reports (DMRs), which are periodic self-reports submitted by facilities to state level environmental agencies (and subsequently to the U.S. EPA) with information about their compliance with permitted effluent limits. We draw upon data from fiscal years 2015-2019 to predict risk of the binary SNC status and the percent exceedance in the first quarter of fiscal year 2020 (October to December 2019; henceforth FY2020-Q1), ultimately developing a training set that reflects features aggregated to be "as of" FY2019-Q4 (or FY2020-Q1 for test). We subset the data to include only permits with reported discharges relevant to the Effluent SNC calculation (discussed more in subsection 4.2.2) and EPA-assigned SNC statuses in the target quarter (FY2020-Q1), thus resulting in a final sample of 40,594 individual and general permits.[1] Second, socio-demographic information on race and median household incomes at the census block group level are extracted from the 2018 version of the five year American Community Survey provided by the U.S. Census Bureau. Third, as a proxy for the degree of environmental burdens communities face, we draw upon figures of population density within three miles of a permittee and the EJcreen percentile flag from the EPA‘s Environmental Compliance and History Online (ECHO) tool. The EJ screen flag, in particular, reflects a combination of demographic as well as environmental data, and consistent with the EPA‘s flag, the indicator we use flags areas at the 80th percentile or above across the US that are suspected to have higher pre-existing potential pollution exposure [14]. Analyses were primarily conducted in R version 4.0 [30], and summary statistics on all variables used in our analyses are featured alongside the data dictionary in Appendix A.2.

[1] Unfortunately 2,200 permittees had incorrect geocoordinates, which prevented us from associating them with demographic information. Therefore the results that feature demographic information reflect a smaller total of 38,394 permittees.

4.2 Risk Prediction Models and Objectives

To examine the distributive implications of noncompliance prediction models, we investigate two different ways of using predictions to generate the priority list of noncompliant facilities to target in FY2020-Q1, simulating how the EPA would employ such risk models using data from the previous quarter (FY2019-Q4). The first method focuses on predicting permit level discharge volumes (regression) and the second focuses on predicting the risk of falling into effluent SNC status (classification). Both models employ a Random Forest model [5] with the same input features; the models only vary in the outcome variable generated. We detail the construction of each outcome variable after a brief discussion of the common elements between the two designs.

4.2.1 Random Forest. Our Random Forest models ingest 27 features drawn from historical discharge volumes, time series predictions of discharge volumes, historical compliance status information over the past two years, and time-invariant permit-level characteristics such as location information, industrial sector, and permit type. Appendix Figure 7 elaborates each input feature and its definition.

4.2.2 Regression: Calculating Permit-Level Overages Across Pollutants and Monitoring Locations. In order to both establish as well as predict the intensity of violations with a continuous measure, we require some aggregation across pollutants. Indeed, even the same permitted pollutant might have a varying allowable limit over time. Building on the nomenclature that EPA uses for monitoring and reporting discharges that result in violations, we first construct an aggregate measure of pollution overages that relies on understanding the volume discharged relative to the permitted limit. Namely, we define our regression objective in terms of the exceedance percentage p, that is, the ratio of the exceedance amount relative to the corresponding limit values.

Each pollution parameter that contributes to the SNC status calculation falls into one of two categories. Category 1 is comprised of so-called 'conventional pollutants' such as Nitrogen and Phosphorous, and Category 2 is comprised of toxic materials and metals. Practically, these two categories set two different thresholds for the percentage over the permitted limit that will trigger an SNC violation - 40% for Category 1 and 20% for Category 2. Since the threshold differs across these groups yet we still seek to develop an aggregate measure across all SNC-eligible pollution parameters monitored in a given permit, we construct a measure based on the pollution parameter-specific percentage thresholds that can trigger the Effluent SNC status. Where we define E as total permit overage, p_i as the percent overage for each category 1 discharge and p_j the percent overage for each category 2 discharge, we generate a composite permit exceedance value that can be represented as the weighted sum of all recorded exceedance percentages of each parameter for a given permit:

$$E = \sum_i \frac{p_i}{40} + \sum_j \frac{p_j}{20} \tag{1}$$

The final weighted sum of all exceedance percentages E for each permit then serves as the outcome variable in the Random Forest Regression model.

4.2.3 Classification: Constructing Synthetic Effluent SNC Status. For ease of exposition and interpretation, we use a simplified effluent SNC status definition that roughly approximates the federal guidelines for the formal effluent SNC calculation [33].

More specifically, we construct a synthetic effluent SNC flag for permittees that aligns with the effluent exceedance conditions EPA uses to determine SNC. The flag is applied under the following two conditions. First, if a permittee discharges beyond its permitted levels by 40% or 20% for any two category 1 and 2 parameters, respectively, then the permittee triggers a "serious" SNC violation. Second, if a permittee has four or more effluent discharges that exceed their limits in any amount over in the past two quarters, they trigger the "chronic" SNC violation flag.

More formally, we can represent these two SNC violation status triggers as follows. First, where $I_{serious}(x)$ is the indicator function for serious violations based on observed parameter exceedance values p_c and thresholds $p_c^* \in \{40, 20\}$ of each category c within a given permit, this flag can be determined as:

$$I_{serious}(p) = \begin{cases} 1 & \text{if } p_c \geq p_c^* \quad \forall \quad c \in \{1, 2\} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Where $I_{chronic}(x)$ is the indicator function for chronic violations, the second flag is determined as:

$$I_{chronic}(p) = \begin{cases} 1 & \text{if } p_c \geq 0 \quad \forall \quad c \in \{1, 2\} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

For each permit, if there is any pollutant q with $\sum_{n=1}^{m} I_{serious}(p_n) \geq 2$ or $\sum_{n=1}^{m} I_{chronic}(p_n) \geq 4$, where m is the number of measurements for q across all monitoring locations in the past two quarters, then the permit will be flagged as being in Effluent SNC. These constructed, permit-level Effluent SNC flags and the probability of a given permit in being flagged with this label serve as the outcome variable in the Random Forest classification model.

This synthetic effluent SNC flag approximates the more complex SNC function reasonably well, as indicated by the 92% of overlapping classifications we determined from our measure compared against the EPA's records.[2]

4.3 Simulated Risk-Based Permit Selection

Based on the outputs of each model, we select facilities to 'target' for compliance efforts as follows. In the classification approach, we use the probability of being in the SNC status as the risk score to rank all permits. We then select the top 50% of those permits, following the NCI objective to halve the SNC rate. In the regression approach, we use the predicted, weighted sums of exceedance percentages to rank all permits. We then select the same number of permittees from the top fraction of the rank-ordered list. This risk-selection procedure simulates actual deployment possibilities under the NCI. The federal EPA and state partners are exploring a range of risk-targeted interventions (e.g., notifications, compliance advisories, inspections) that would shift enforcement resources toward such risk-scored facilities with the goal of improving environmental

[2] We note that not all Effluent Violations that meet the Chronic or TRC threshold are identified as Effluent SNC under the official EPA flag. For example, a permit may also independently trigger a permit compliance schedule event violation that supersedes an Effluent SNC in EPA's SNC categorization hierarchy. In addition, not all effluent violations are eligible for detection as SNC, with the full set of criteria documented in [33].

protection. Our aim is thus to study what the distributive impact of regression vs. classification approaches might be for a set of facilities targeted in such interventions.

4.4 Evaluating Distributive Impacts

To evaluate distributive impacts, we first link each permit back to facility-level information indicating in which census block group (CBG) each permit is located. Next, we associate each permittee with demographic data corresponding to their CBG, as made available in the ACS data or, in the case of the population density information, as extracted from ECHO. We then evaluate the distributions of features associated with the targeted facilities in each targeting protocol using two-sided t-tests (for means), Kolmogorov–Smirnov (ks) tests (for distributions), and the Wilcoxon ranked sum test (for medians).[3] Finally, we graphically represent the differences between select demographics of the targeted (or risk-selected) permittees in quantile-quantile (QQ) plots.

5 RESULTS

5.1 General vs. Individual Permits

Although taking the step to expand the purview of compliance initiatives to include minors for the first time marks an ambitious seven-fold expansion in the number of facilities under such an initiative, we now consider how delaying the inclusion of the approximately 37,000 general facilities in the National Compliance Initiative can shape the distribution of environmental compliance resources.

Figure 3 demonstrates the gaps created by focusing the NCI only on individual permits. To illustrate the variation in state-level permitting decisions even for the same type of facility, Figure 3 focuses on a subset of wastewater treatment facilities across the United States, namely those that manage sewage (as opposed to industrial effluent). The top panel maps these sewage-handling wastewater treatment plants across the US, colored by whether they have general (orange) or individual permits (blue), and the bottom panel zooms in on the neighboring states of Virginia and North Carolina, illustrating seemingly arbitrary differences in state permitting schemes.

Furthermore, Figure 4 suggests that the variation in individual versus general designation are not necessarily driven by differences in the amounts of effluent characteristics that are permitted under each type.[4] Figure 4 plots the distribution of effluent limits, i.e., the maximum permitted value for each discharge parameter, for six common water discharge parameters across the 10,950 general and 15,020 individual permittees shown in Figure 3. The substantial overlapping regions between the effluent limits suggests there is significant overlap between the effluent characteristics of these two types of facilities. Last, the exclusion of general permits from the NCI is not necessarily because general permittees are always in compliance. Table 1 presents the compliance statistics for general and individual permits, calculating SNC as if the NCI were applied to all facilities. Roughly 10.5% of individual permits overall – that is, across major and minor permits – are in effluent SNC, compared to 3% of general permits, although that difference is largely a function of higher non-reporting among general permits. The overall SNC rate between the two permit types is comparable. In short, due to discretionary differences between state permitting protocols, the NCI excludes from its purview a large number of functionally similar facilities.

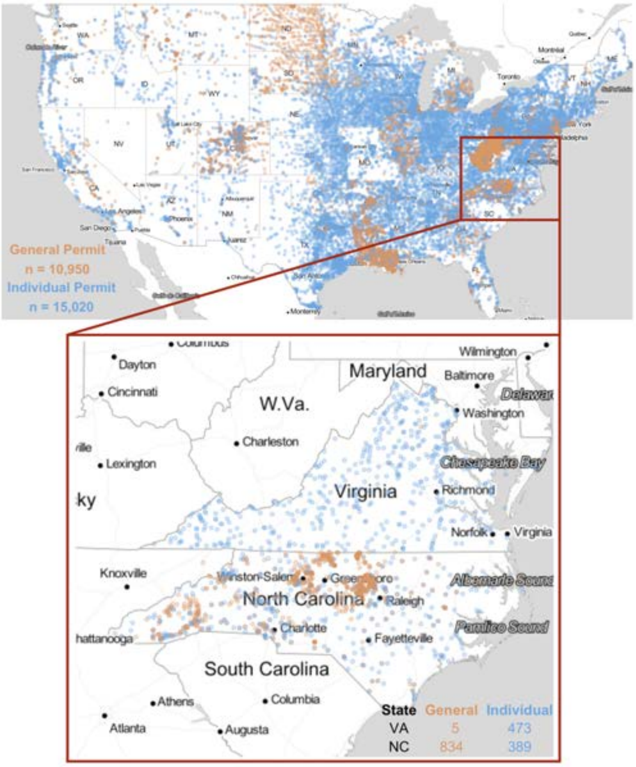


Figure 3: (a) As of FY2018, 42% of wastewater treatment facilities across the US responsible for handling sewage have general permits rather than individual permits – and therefore do not count for the National Compliance Initiative. (b) How wastewater treatment facilities are classified varies within and across states, as seen by the permit classification for Wastewater Treatment Plants that handle sewage in the two neighboring states of Virginia and North Carolina.

What impact does this policy decision have? Table 2 provides some descriptive information to understand the impacts of first expanding the scope of the NCI to include individual permits, and second, to delay the inclusion of general facilities. First, based on eligible facilities, we can see that including individual minor facilities shifted the mass of attention towards lower density areas with fewer minority populations, on average. Second, general permits exceed the total number of individual permits by nearly a third. Third, both the general and individual categories have large numbers of wastewater treatment plants, which highlights that very

[3] This test is also known as the Mann-Whitney test.

[4] We acknowledge there are legitimate reasons distinguishing the issuance of individual versus general permits. General permits are intended to be easier, cheaper, faster, as they were designed for more standardized types of operations that, in principle, should require less scrutiny. Where used, individual permits typically build off the guidelines set in general permits but are then customized to the types of effluent that a individual permittee may have in addition to the effluent common among standard general permits.

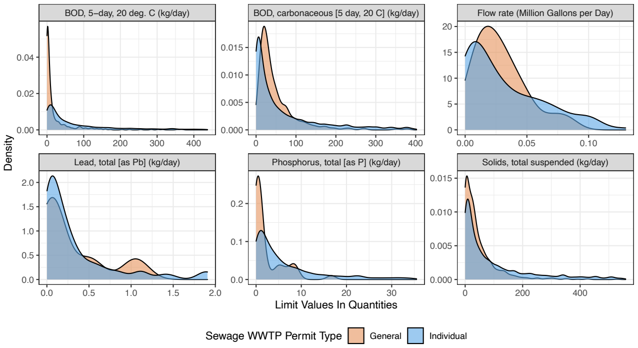


Figure 4: Discharge limits for sewage-handling wastewater treatment plants (WWTP) that handle sewage but are permitted under the general (n=10,950) or individual (n=15,020) category are broadly similar across a series of six commonly reported discharge parameters.

	General	Indiv. Major	Indiv. Minor
Number of Facilities	37,804	3,619	24,682
Effluent SNC	953	503	2,480
Facility Type*			
Wastewater	16,358	3,398	23,488
Stormwater	20,034	124	738
Major	30	3,619	0
Demographics**			
Avg Population Density	1,173	1,298	663
Avg Median HH Income	60,673	59,040	58,838
Avg Percent Minority	24	30	17

Table 2: Characteristics of General, Individual Major, and Individual Minor Permits *Categories not exclusive **All differences statistically significant (p <0.001) between general and individual permits and for population density and percent minority between the two types of individual permits.

similar facilities can be subject or exempted from the NCI, solely due to permitting vagaries.[5] Third, and most importantly, general permittees tend to be located in denser areas with a higher share of minority individuals, relative to the average individual permit included in the scope of the original NCI. This shows that the design decision to exclude general permittees from the NCI itself had disparate impact, shifting environmental remediation efforts toward non-minority regions. Last, while one rationale for excluding general permittees might be that their effluent SNC rate is much lower (3%), as stated above, the total SNC rate is comparable due to high rates of failures to submit.[6] Such non-submissions are themselves subject to penalties under the Clean Water Act, reflecting the importance of information reporting under the Act.

5.2 Oracle Test

As a first benchmark, we consider the 'oracle' test, assuming that the decision maker is omniscient about effluent exceedances and SNC status in the target quarter FY2020-Q1. As described above, we select the top riskiest half of permittees expected to be in SNC (n = 1,392) on the basis of that full information and compare the distributive effects using a selection rule that focuses on top exceedances versus SNC status. This comparison mimics the regression approach (which focuses on top polluters regardless of SNC status) and the classification approach (which focuses on SNC status regardless of the level of pollution). To understand the distributive effects of choosing one approach over the other, the left panels of Figure 6 depict QQ plots, comparing quantiles of attributes from risk-selected facilities based on classification (x-axis) and regression (y-axis). Identical selections would line up along the 45-degree line. Instead, we observe substantial evidence of the potential for disparate impact based on this oracle test. The regression model selects a subset of permittees with higher shares of minority populations located in more densely populated areas than the classification counterpart (top and bottom left panels). These distributional shifts are statistically significant (p-value < 0.01 based on all three tests for both features). The regression model also focuses on a higher fraction of areas flagged as "vulnerable"[7] (18.7% versus 13.5%), in effect suggesting that a higher share of communities in which these regression-selected facilities are located already have a series of preexisting exposures and vulnerabilities.

The oracle test illustrates the potential for disparate impact based on the choice of regression versus classification, but the impact of model-based inferences is less clear. Depending on the model, actual risk models may amplify or attenuate the distributive impact identified in the oracle.

5.3 Performance Assessment

We provide brief performance statistics of the RF regression and classification models here. Figure 5 provides a calibration plot for the classification model in left panel, binned by deciles. As expected, the classification model is properly calibrated, with bins falling along the 45-degree line. The classification model has an Area Under the (ROC) Curve (AUC) of 0.93 and an AUC on the precision-recall (PR) curve of 0.78. The right panel plots deciles of predicted effluent exceedances on the x-axis against the difference between observed and predicted exceedances on the y-axis from the RF regression model. The regression predictions perform well along all but the most extreme bin, which is driven by a small number of extreme outliers. Including all data points results in an RMSE of 240, compared to a mean imputation baseline of 288. Once omitting the top 3% of outliers, however, the RMSE drops to 117.

As a sanity check, we examine how facilities that were risk-selected either via regression or classification perform in the test period on SNC and exceedance (or overage) percentage. Table 3

[5] General permits include far more stormwater permits, which, for instance, regulate stormwater runoff from construction and industrial activities. The dominance of stormwater permits also suggests part of the challenge of effectively monitoring and managing their violations: while some preventative measures may be taken to reduce violations, stormwater SNCs may stem, at least in part, from stochastic weather events.

[6] An effluent violation can only be ascertained if the permittee submits the DMR. Thus, the 3% effluent SNC rate is a function of the high rates of failure to submit. The "true" effluent SNC rate would likely be much higher in the counterfactual where all general permittees submitted their DMRs.

[7] That is, they are flagged as in the 80th percentile or above of the EJScreen Monitoring Tool.

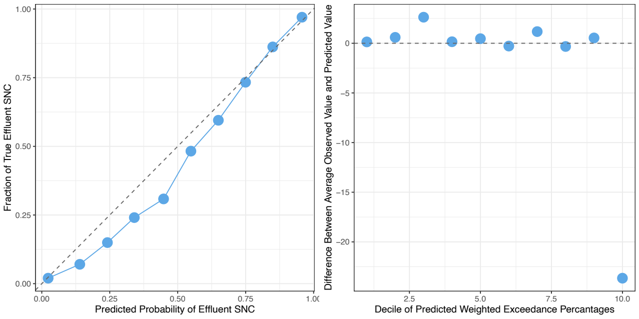


Figure 5: Calibration plot for classification model and binned difference between observed and predicted values for the regression model.

shows that, on average, the regression model identifies more permittees with large overages (aggregated over all parameters), whereas the classification model identifies more permittees flagged as with the effluent SNC status. Comparing the risk-selected samples also reveals that the models agreed for roughly half of the selected facilities. The remaining differences between what we observe in these two protocols are thus all attributable to the characteristics of the remaining facilities. Appendix B provides more detail on the workflow and other performance statistics.

Model	Overage %	SNC Status
Regression	210	87%
Classification	116	97%

Table 3: Comparing Model Performance on Substantive Measures: The Regression model identifies more permittees with large predicted and actual overages, aggregated over all parameters, and the classification model identifies more permittees with the effluent SNC status flag.

5.4 Disparate Impact of Model-Based Selections

We now return to the measures of disparate impact and add these to compare against the oracle test in Figure 6. The right column presents the distributive shift for proportion minority, income, and population density from classification (x-axis) vs. regression (y-axis). As before, these represent the risk-selected facilities under either model. Under no distributional shift, the QQ plot should line up along the 45-degree line.

Instead, we observe statistically significant differences in the means, medians, and distributions of the percent minority and the population density measures. The extent of the differences amounts to, on average, a 2% difference (p-value < 0.01) in the share of communities with minority populations targeted under the regression-focused approach, going from approximately 16.7% in the classification up to 19% under regression. The QQ plots further reveal that the shift appears to be uniform across the distribution. We observe no statistically distinguishable differences between the median household income of the permittees targeted under regression versus classification, with both median household incomes resting around $57-58k (SD = 25k). The final panel reveals that classification directs relatively more attention to permittees in areas with about 400 individuals per square mile within a 3 mile radius of the permittee, with a notable set of permittees with 10 people per square mile or below, whereas regression focuses on permittees in areas with population densities that are, on average, nearly 700 people per square mile and above. Importantly, regression prioritizes several facilities in some of the densest places in the United States, including nine permittees located New York, Massachusetts, and California that feature over 10,000 people per square mile within a 3 mile radius of those facilities.

These results show that targeting environmental protection based on the intensity of pollution exceedances would focus on areas with higher shares of minority populations in denser, more urban areas.

6 CONCLUSION

In this paper, we have drawn upon data on Clean Water Act permits, historical pollution discharge and compliance records, institutional knowledge of regulatory implementation details from extensive engagement with federal EPA, and census data to demonstrate how algorithmic design for environmental enforcement can identify levers that can exacerbate or mitigate disparate impact. We simulate which sets of permittees would be targeted if EPA seeks to focus on those with the highest emission overages (regression) compared with selecting permittees the highest probability of falling into the SNC status (classification), and we show how an objective that focuses on exceedance intensity would redirect compliance efforts away from more rural, smaller facilities and towards permittees situated in more densely populated environments with higher shares of minority individuals. This finding holds with both observed values (the oracle test) as well as the ML-based risk assessment models to enable prospective interventions that prevent and reduce significant noncompliance.

While we believe that this work adds an important case study of algorithmic fairness "in the wild," so to speak, we also acknowledge a few important limitations. First, mirroring the definition of SNC, our models implicitly assume that every additional percent exceedance can be compared with another percent exceedance of the same category. A facility with a 100 unit limit of nitrogen that discharges 150 units counts equally towards the SNC status trigger and exceedance percent estimation as a facility with a 1000 unit limit of phosphorous that discharges 1500 units. Absent other mechanisms to put these different effluent types and amounts on a common scale of harm, we in effect assume that pollution limits were established with solid knowledge about which amounts pollution would result in social and environmental damages in a given area.

We recognize that the extent of harm from additional pollution may be nonlinear and vary considerably with baseline levels of water quality impairment and vulnerability across regions. To capture some measure of vulnerability, we include the EJScreen flag into our assessment of distributive implications. Even though the EJScreen flag is a coarse instrument, it attempts to measure areas likely face a high degree of pre-existing cumulative exposures and risks. Nonetheless, the goal of translating the coarse measure of

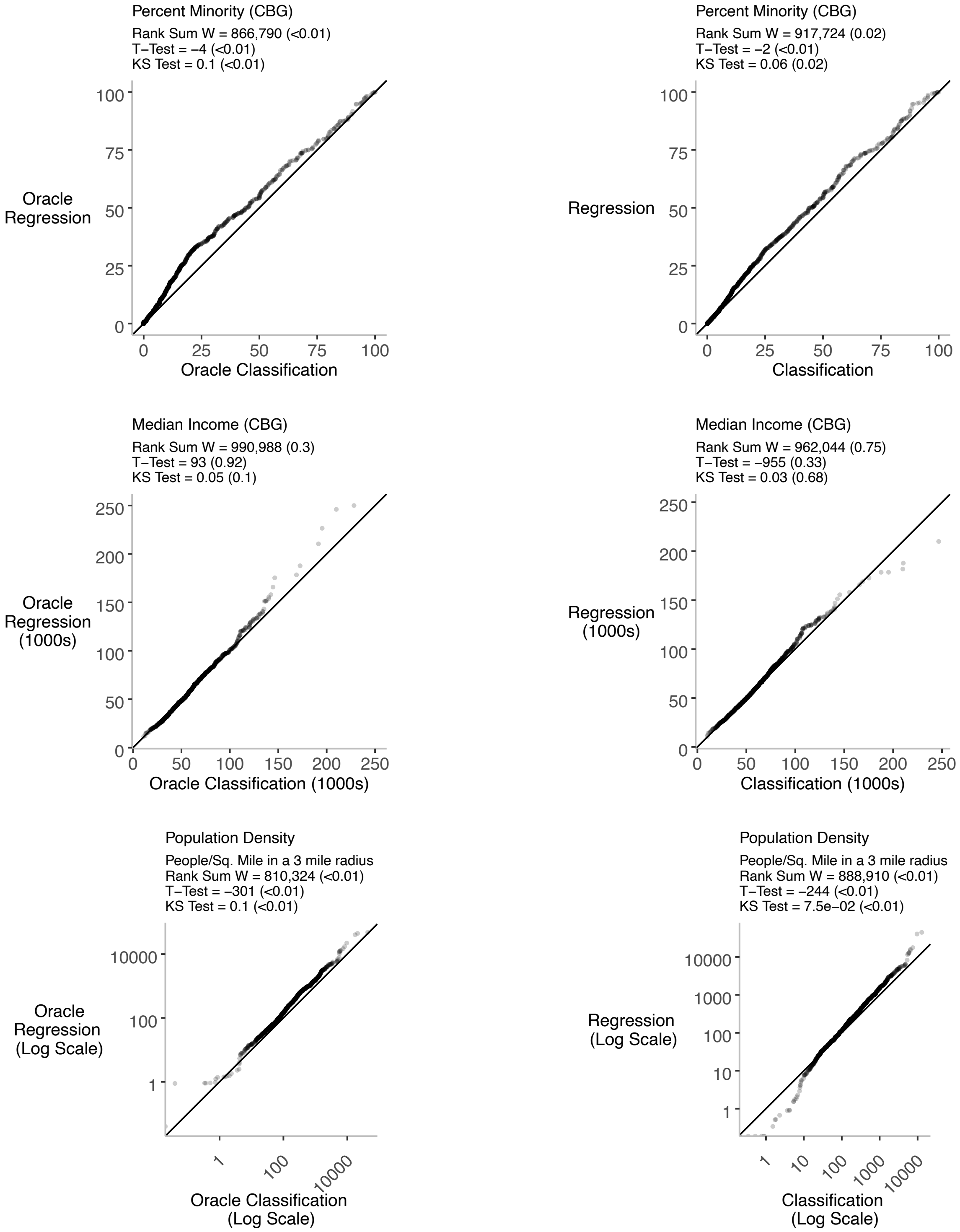


Figure 6: The quantile-quantile (QQ) plots above reflect the differences in the demographics of the areas in which the 1,392 NPDES permittees flagged under each targeting protocol are located. The left column reflects which facilities would be targeted if full effluent information was known for FY2020-Q1, and the right shows the demographics surrounding permittees would be targeted if prioritization stemmed from predictions for FY2020-Q1. p-values are presented in parentheses for rank sum, t, and KS tests.

effluent percent exceedances into the social and environmental harms of the context they are discharged into, is an important, if challenging, open area of research.

Second, while we select facilities based on these coarse risk or exceedance estimates, we do not have insight into the types of enforcement actions based on this prioritization and their causal effects. Compliance efforts can run the gamut from inexpensive informational interventions to careful but costly on-the-ground facility inspections and monitoring. The extent to which this list translates into meaningful changes in conditions on the ground may rest heavily on the extent to which the discharge in question poses harms to the communities exposed to it as well as the feasibility and costs of remediating the noncompliance.

Notwithstanding these limitations, the core of our argument is essentially thus: the use of seemingly simple and clear objectives, such as the 50% reduction in SNC, can mask important policy decisions. In selecting the SNC rate among individual permittees as the measure for evaluating the performance of a signature national compliance initiative, the EPA implicitly makes a decision about whose compliance is important, what types of violations should be treated equivalently, and what types of compliance efforts should be encouraged. Although we cannot claim that every additional unit of pollution has an equal impact (due to changes in distance from population centers, changes in concentration stemming from dilution as well as mixing of pollution once in contact with waterways, etc.),[8] our investigation has clarified a basic trade-off. Conventionally, whether to take an enforcement action against the facilities that are barely out of compliance or facilities that are seriously out of compliance might involve a calculus of the resource cost. Barely noncompliant facilities, for example, might be cheaper to get back into compliance. Algorithmic design shows that there is another dimension: the largest polluters are also more likely to reside in vulnerable and disproportionately minority communities. By focusing the NCI on the violation rate among individual permits, the NCI bypasses a potentially important mechanism for reducing disproportionate harms.

Third, although we focus on effluent-related SNC in this paper, the NCI as a whole covers the full range of reasons why a facility may trigger that label. Specifically, as presaged by Table 1, over half of permits in SNC fall into this category due to nonreporting of their DMRs. And if these nonreporting facilities are already substantively in compliance with their permits, the NCI could then potentially be achieved without actually reducing pollution amounts. (In actuality, many believe nonreporting tends to mask a variety of other problems, including permit exceedances.) Accordingly, our approach may well underestimate the distributive implications that the overall SNC-focused NCI has compared to a specific discharge-over-limit reduction goal focused on reducing harms in the places where they matter most.

Last, while our models approximate what government agencies are deploying [12], the models could be improved in a number of respects. The RF regression, for instance, does not perform well for some of the extreme outliers we observe in the data; forms of robust regression would help to reduce the influence of such data points. While we have tried one form of dimensionality reduction with ARIMA, sequence-based models may enable better utilization of the underlying panel time series structure of the monthly DMR data. Nonetheless, the application illustrates some of the key challenges in machine learning with the complexities of real world government data and application.

To conclude, as government agencies expand the use of algorithmic decision-making in guiding and executing policy decisions, the many micro policy and engineering choices on which the top-line objectives are implemented can themselves generate unintended impacts. Our research has added a case study in an important area of environmental sustainability and regulatory enforcement, where algorithmic and policy design may be inextricably intertwined. Algorithmic fairness in the regulatory state may require grappling with the policy objectives and design themselves, but can also, if carefully done, shed light on the impacts of prior policy choices. Last, this case study illustrates how academic-agency collaborations can ensure greater attention to identifying and mitigating disparate impact in algorithmic decision making.

ACKNOWLEDGMENTS

We thank the Stanford Woods Institute for the Environment and Schmidt Futures for funding support, Dan Palmer for helpful feedback, and Andrew Schulman, John Veresh, and Beth Ragnauth for explaining and helping interpret EPA databases.

REFERENCES

[1] Spencer Banzhaf, Lala Ma, and Christopher Timmins. 2019. Environmental justice: Establishing causal relationships. *Annual Review of Resource Economics* 11 (2019), 377–398.

[2] Spencer Banzhaf, Lala Ma, and Christopher Timmins. 2019. Environmental justice: The economics of race, place, and pollution. *Journal of Economic Perspectives* 33, 1 (2019), 185–208.

[3] Susan Bodine. 2018. *Transition from National Enforcement Initiatives to National Compliance Initiatives.* Government Memo.

[4] George.E.P. Box and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control.* Holden-Day, Hoboken, New Jersey.

[5] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency.* PMLR, New York, NY, USA, 134–148.

[7] Danielle Keats Citron. 2007. Technological due process. *Washington University Law Review* 85 (2007), 1249.

[8] Bill Clinton. 2014. Executive Order 12898 of February 11, 1994: Federal Actions To Address Environmental Justice in Minority Populations and Low-Income Populations. https://www.archives.gov/files/federal-register/executive-orders/pdf/12898.pdf

[9] Code of Federal Regulations. 2007. Title 40 - Protection of Environment: Section 123.45 - Noncompliance and program reporting by the Director. https://www.govinfo.gov/content/pkg/CFR-2007-title40-vol21/xml/CFR-2007-title40-vol21-sec123-45.xml (Accessed on 09/21/2020).

[10] Cary Coglianese and David Lehr. 2016. Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal* 105 (2016), 1147.

[11] David Freeman Engstrom and Daniel E Ho. 2020. Algorithmic Accountability in the Administrative State. *Yale Journal on Regulation* 37 (2020), 800–54.

[12] David Freeman Engstrom, Daniel E. Ho, Catherine Sharkey, and Mariano-Florentino Cuéllar. 2020. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies.* Administrative Conference of the United States, Washington DC, United States.

[13] EPA. 2015. National Pollutant Discharge Elimination System (NPDES) Electronic Reporting Rule; Final Rule. https://www.gpo.gov/fdsys/pkg/FR-2015-10-22/pdf/2015-24954.pdf

[14] EPA. 2019. *EJSCREEN Technical Documentation: environmental justice screening and mapping tool.* Technical Documentation. EPA. https://www.epa.gov/sites/production/files/2017-09/documents/2017_ejscreen_technical_document.pdf (Accessed on 10/01/2020).

[8] For more on these challenges, see Olmstead.

[15] EPA. 2020. Detailed Facility Report Data Dictionary. https://echo.epa.gov/help/reports/dfr-data-dictionary

[16] EPA. 2020. ICIS NPDES Download Summary. https://echo.epa.gov/tools/data-downloads/icis-npdes-download-summary

[17] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, New York, New York.

[18] Jeffrey M Gaba. 2007. Generally illegal: NPDES general permits under the Clean Water Act. *Harv. Envtl. L. Rev.* 31 (2007), 409.

[19] Cynthia Giles. 2020. *Next Generation Compliance: Environmental Regulation for the Modern Era.* Harvard Law School Environmental and Energy Law Program, Cambridge, Massachusetts.

[20] Cassandra Handan-Nader and Daniel E Ho. 2019. Deep learning to map concentrated animal feeding operations. *Nature Sustainability* 2, 4 (2019), 298–306.

[21] David A Hindin and Jon D Silberman. 2016. Designing More Effective Rules and Permits. *Geo. Wash. J. Energy & Envtl. L.* 7 (2016), 103.

[22] M Hino, E Benami, and N Brooks. 2018. Machine learning for environmental monitoring. *Nature Sustainability* 1, 10 (2018), 583.

[23] Rob J. Hyndman and Yeasmin Khandakar. 2008. Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software* 27 (2008), 1–22. Issue 3.

[24] Pauline T Kim. 2017. Auditing algorithms for discrimination. *University of Pennsylvania Law Review Online* 166 (2017), 189.

[25] David M Konisky. 2009. Inequities in enforcement? Environmental justice and government performance. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 28, 1 (2009), 102–121.

[26] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *University of Pennsylvania Law Review.* 165 (2016), 633.

[27] Kelly Maguire and Glenn Sheriff. 2011. Comparing distributions of environmental outcomes for regulatory environmental justice analysis. *International journal of environmental research and public health* 8, 5 (2011), 1707–1726.

[28] Sheila M Olmstead. 2010. The economics of water quality. *Review of Environmental Economics and Policy* 4, 1 (2010), 44–62.

[29] Irina Pencheva, Marc Esteve, and Slava Jankin Mikhaylov. 2020. Big Data and AI – A transformational shift for government: So, what next for research? *Public Policy and Administration* 35, 1 (2020), 24–44.

[30] R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ Version 4.0.

[31] United States Code. 1977. Federal Water Pollution Control Act, As Amended by the Clean Water Act of 1977. https://www3.epa.gov/npdes/pubs/cwatxt.txt (Accessed on 09/20/2020).

[32] US Environmental Protection Agency Environment and Compliance History Online. 2020. Data Downloads. https://echo.epa.gov/tools/data-downloads#downloads (Accessed on 09/28/2020).

[33] US Environmental Protection Agency Integrated Compliance Information System National Pollutant Discharge Elimination System. 2018. Technical Specification Document: RNC Processing Technical Design, Version 1.19. https://icis.zendesk.com/hc/en-us/articles/207065796-ICIS-Web-Technical-Specifications-Program-Reports-RAD-WebRIT-Reissuance-Related-Activities-Reports-Universes-RNC- (Accessed on 09/28/2020).

[34] US Environmental Protection Agency Office of Enforcement and Compliance Assurance. 1995. Revision of NPDES Significant Noncompliance (SNC) Criteria to Address Violations of Non-Monthly Average Limits. https://www.epa.gov/sites/production/files/documents/revisnpdessnc.pdf (Accessed on 09/21/2020).

[35] US Environmental Protection Agency Office of Enforcement and Compliance Assurance. 2019. National Compliance Initiative: Reducing Significant Non-Compliance with National Pollutant Discharge Elimination System (NPDES) Permits. https://www.epa.gov/enforcement/national-compliance-initiative-reducing-significant-non-compliance-national-pollutant

[36] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications* 11, 1 (2020), 1–10.

[37] Kristoffer Wikstrom, Trisha Miller, Heather E Campbell, and Michael Tschudi. 2019. Environmental inequities and water policy during a drought: Burdened communities, minority residents, and cutback assignments. *Review of Policy Research* 36, 1 (2019), 4–27.

[38] Bernd W. Wirtz, Jan C. Weyerer, and Carolin Geyer. 2019. Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration* 42, 7 (2019), 596–615.

APPENDIX

A ELABORATED DATA DESCRIPTIONS

A.1 Code Availability

The code used in this project is available at https://github.com/reglab/snc-distributive

A.2 Data Dictionary

(1) **Individual Permits**: An Individual NPDES Permit is a permit specifically tailored to an given facility that discharges effluent into US waters.
(2) **General Permits**: A general permit covers a group of dischargers that, in principle, should have similar characteristics within a given geographical location.
(3) **Major Facilities**: There are two types of major facilities, municipal and industrial. Major municipal dischargers include all facilities with design flows greater than one million gallons per day and facilities with approved industrial pretreatment programs. Major industrial facilities are determined based on specific ratings criteria developed by US EPA and/or the states.
(4) **Minor Facilities**: A minor facility is a discharger with a design flow of less than one million gallons per day (MGD) that has not been determined to have an actual or potential adverse environmental impact that would classify the discharger as major.
(5) **Stormwater Permits**: the NPDES Stormwater program regulates stormwater discharges from three potential sources: municipal separate storm sewer systems, construction activities, and industrial activities. We identified stormwater permits based on the NPDES Program Areas outline in the permit data.
(6) **Wastewater Permits**: NPDES permits establish discharge limits and conditions for discharges from municipal wastewater treatment facilities to waters of the United States. We identified which facilities are wastewater permits based on fields that indicate that a facility is either a Publicly Owned Treatment Works (POTW) and/or is classified as a wastewater permit under the NPDES Program Areas field. We identified Wastewater Treatment Plants that handle sewage by using the Industrial Classification Codes (NAICS or SIC codes) associated with the facility. NAICS Code 221320 indicates Sewage Treatment Facilities; SIC Code 4952 indicates Sewerage Systems.
(7) **Median Household Income**: Median Household Income as determined by the American Community Survey of the Census Block Group in which the facility is located.
(8) **Percentage Minority**: Percentage of population within a 3-mile radius of the facility that is non-white.
(9) **Population Density**: Persons per square mile in a 3-mile radius of the facility.
(10) **EJ Screen**: The EJ (Environmental Justice) Screen was developed by the EPA to assess the potential for disproportionate environmental impacts and other significant environmental justice concerns for populations across the country. Each EJ index is a combination of environmental and demographic information. The EJ Flag as used in the paper reflects facilities that rank at the eightieth percentile or above in the EJScreen distribution.
(11) **Effluent Significant Non-Compliance**: A facility is considered to be in effluent SNC if, for the same pollutant parameter, either: (A) the facility exceeds the discharge limit (by any amount) in at least 4 of the preceding 6 months (referred to as "Chronic" Violation) or (B) the facility triggers a "Technical Review Criteria" (TRC) Violation (see definition for Group 1 and 2 Pollutants) in at least 2 of the preceding 6 months.
(12) **Exceedance Percentage**: The amount by which the facility exceeded the limit value for each parameter, outfall, and monitoring location.
(13) **Group 1 and 2 Pollutants**: The Code of Federal Regulations [9] define two main groups of pollutants with respect to the calculation of effluent significant non-compliance. Group 1 pollutants, which are conventional pollutants (e.g., Nitrogen, Phosphorous, total suspended solids, detergents, oils, and total organic carbon), are subject to a 40% threshold beyond the permitted limit above which an exceedance would trigger a TRC violation. For Group 2 pollutants, which are toxic pollutants (e.g., most metals, cyanide, and toxic organic compounds), that threshold is 20%.
(14) **Historical Quarterly Non Compliance Report (QNCR) Statuses**: The historical compliance status for each facility at the quarterly level as indicated by the publicly available data downloaded from [16].
(15) **Limit**: The specified discharge allowance described in the NPDES permit for each facility, outfall, monitoring location, and pollution parameter.
(16) **Significant non-compliance (SNC)**: The SNC designation refers to the most serious class of Clean Water Act violations considered to pose a threat to U.S. waters and/or public health. There are two main categories of SNC: Non-Reporting and Effluent. The remaining types of SNC most frequently refer to violations of an agreed upon compliance schedule.
(17) **Statistical Base**: For each limit, there is a set of defined statistical analyses to be used for the limit value. Examples include: arithmetic mean, geometric mean, median, etc. Furthermore, as outlined in a 1995 EPA legal memorandum [34], there is a distinction between monthly vs non-monthly limits. Non-monthly limits refer to limits written with any other timescale other than monthly (e.g. annual, daily, etc.)
(18) **Total Design Flow Number**: The flow that a permitted facility was designed to accommodate, expressed as millions of gallons per day (MGD).

A.3 Summary Statistics

Our models draw upon 27 features constructed from three main sources of data: permit-level metadata (9 features), two years of historical compliance statuses (8 features), and time series data from self-reported DMRs (10 features). While DMRs represent very rich data, dimensionality quickly explodes relative to the number of facilities, as each facility may have hundreds of parameters, measured monthly at distinct discharge points. We hence describe in Section B.1.1 how we develop time series forecasts for the target quarter to reduce dimensionality of this data. Table 4 gives detailed descriptions and summary statistics of each feature used.

B RISK MODEL DESIGN

We now describe the full data and machine learning pipeline. We first discuss the features and the feature engineering steps taken for the model. We then describe the training and scoring procedures. Lastly, we describe how we determine high-risk permits from both the classification and regression approaches. Figure 7 demonstrates the workflow graphically.

B.1 Feature Engineering

B.1.1 Using Time Series Models to Predict Effluent Volumes. To take advantage of the rich information contained in the time series data of the monitoring reports, we train auto-regressive integrated moving average (ARIMA) models to forecast discharge volumes. We then aggregated the pollutant-parameter level forecasts to the permit level as described in Section 4.2.2. Based on the definition of effluent SNC, two quarters of data are needed to determine the SNC status. Simulating the scenario where the EPA is making a resource allocation and intervention decision at the end of FY2019-Q4 about what do to for FY2020-Q1, we include the aggregated features for FY2019-Q4 using true historical values (because we would have already observed the data) and for FY2020-Q1 using predicted effluent values (because we are predicting into the future) as inputs to the Random Forest Model.

ARIMA utilizes past values, lags, and lagged forecast errors to forecast future values, given an order of differencing d to make the time series stationary [4]. The general model used in our analysis can be represented as:

$$\hat{Y_t} = \alpha+\beta_1 Y_{t-1}+\beta_2 Y_{t-2}+...+\beta_p Y_{t-p}\epsilon_t+\phi_1\epsilon_{t-1}+\phi_2\epsilon_{t-2}+...+\phi_q\epsilon_{t-q} \tag{4}$$

where $\hat{Y_t}$ reflects the discharge volume of a given permit parameter for the target quarter FY2020-Q1, and Y_{t-p} is the value in any previous monitoring period p, e.g. Y_{t-1} is the lag 1 of the time series. The error terms ϵ_{t_q} are the errors of the auto-regressive model of the respective lags Y_{t-p}, where q refers to the number of lagged forecast errors that go into the ARIMA model. α is the per-permit constant, β is the auto-regressive coefficient, and ϕ is the moving average coefficient. An ARIMA model is then characterized by three terms - d the order of differencing, p the number of Y to be used predictors, and q the number of lagged forecast errors.

We implement the ARIMA model using the *forecast* package in R [23, 30], which enables automatic selection of hyperparameters based on measures of in-sample errors (AIC, BIC, or AICC). In our case, each permit parameter series is treated as an independent univariate time series and we use different ARIMA models (determined by the three characteristics d, p, and q) for all possible time series. Figure 8 provides examples of the signal extracted from these time series models. The left panel plots an instance of a parameter forecast to be likely to exceed the limit (dashed line). The right panel indicates a facility that has had historical exceedances, but is forecast to be in compliance due to downward trends in exceedances.

B.2 Data Processing and Modeling

To process the data, we first remove all known data errors in the overage percentage column; such data errors are coded as "2,147,483,650" and "99,999" percent [15]. Less than 0.1% of records were removed from the DMRs based on such errors.

We then treat missing values by imputing 0 for all missing numeric features and adding "missing" as an additional level to the categorical features. This can be grounded by the assumption that missing (numeric) values in fact represent 0 (or undetectable) levels of discharges for that pollutant. For categorical variables, the additional missing category enables decision trees to branch on missingness.

To reduce the influence of extremely high values in the regression model, we winsorize the outcome variables, capping the exceedance percentages to be equal to or lower than 99, 999%.

After data processing, we train the Random Forest models on a cross section of FY2019-Q4 data (n = 39,352) and test the models on FY2020-Q1 data (n = 40,594). We use 10-fold cross validation to tune hyper-parameters. For classification, a model with 500 trees and 5 variables tried at each split yields the best performance; for regression, 500 trees and 9 variables yield the best performance.

C RISK MODEL PERFORMANCE

We provide more details on the predictive performance of both the classification model and regression model in this section. We compare the prediction results with the actual SNC statuses and exceedance percentages in FY2020-Q1.

C.1 Classification

Using a 50% threshold to trigger an SNC flag, the classification model achieves a 95% accuracy rate. In the policy setting, where inspection resources are limited, we might want to vary the threshold to prioritize a smaller set of facilities. To evaluate the predictive performance in different thresholds, the left panel in Figure 9 plots the ROC (Receiver Operating Characteristic) curve with 0.93 as the area under curve (AUROC). As detailed in Section B.2, our prediction sample is imbalanced, with 9% of permits belonging to the positive class. To capture the trade off between precision and recall, we plot the precision and recall curve in the right panel of Figure 9, with AUC for the precision-recall (PR) curve reaching 0.78. The PR curve retains high precision with up to 50% recall, but then precision drops significantly.

We use feature importance to assess the relative weight of inputs. If a permit already triggered SNC in the previous quarter and the violations are not resolved, the permit will also be under SNC in the next quarter. This allows the EPA to get a reasonable sense of which permits will be under SNC in the next quarter. For the same reason, the historical status from the prior quarter has high predictive power. The features constructed from FY2020-Q1 ARIMA forecasts also add valuable information to the model. Permit-level meta characteristics add important context to the model, but feature importance is significantly below that of the time varying features.

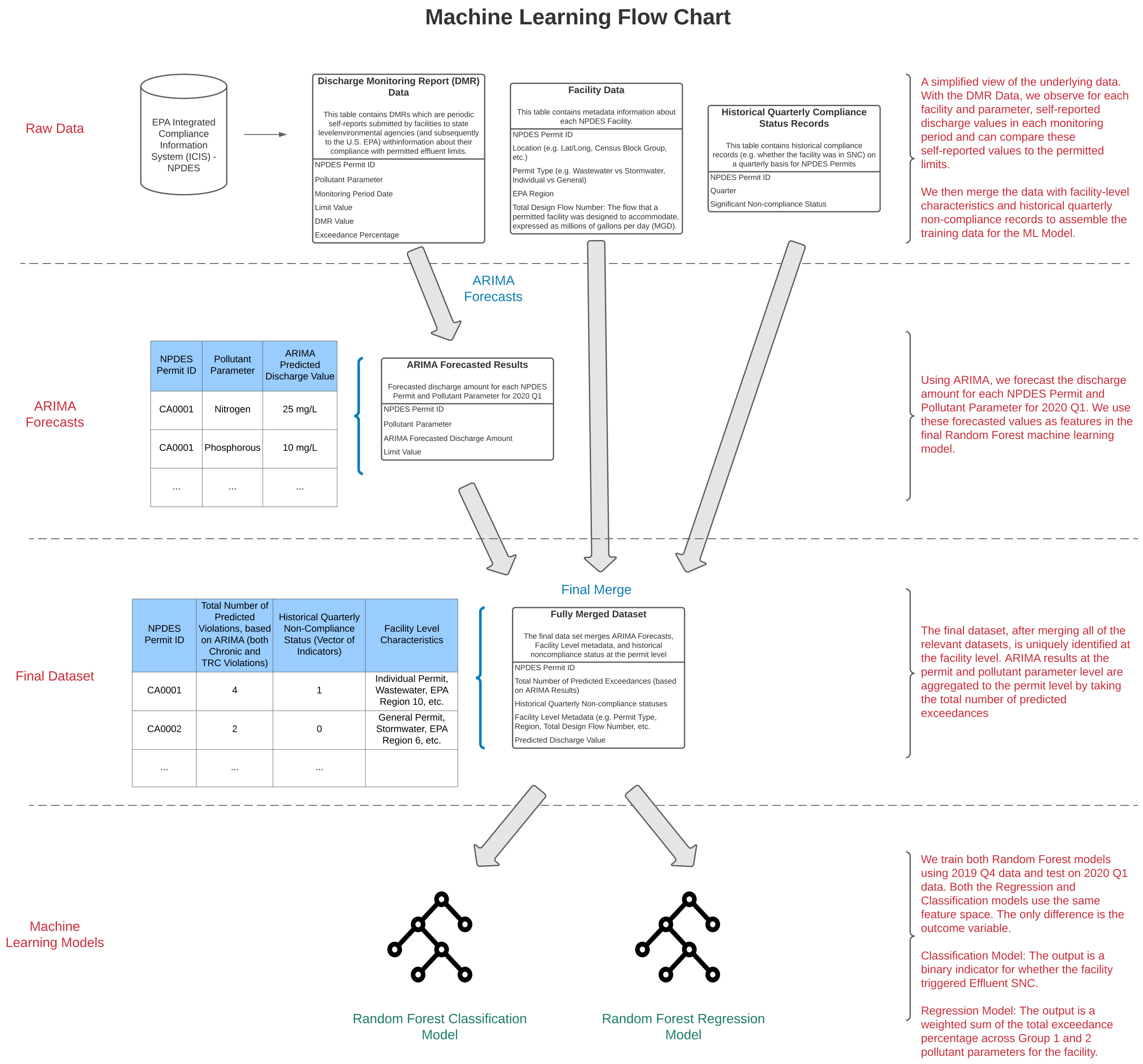
Numerical Features

	Feature	Mean (SD)	Range	Missing Count (%)
	Aggregated Time Series Features From the Previous Quarter*			
1	Measurement Count Across All Parameters	18.25 (22.16)	0 - 995	10,585 (13%)
2	Unweighted Sum of Exceedance Percentages Across All Category 1 and 2 Parameters	407.07 (7,653.49)	0 - 999,980	10,585 (13%)
3	Weighted Sum of Exceedance Percentages Across All Category 1 and 2 Parameters	14.28 (273.23)	0 - 27,010.43	10,585 (13%)
4	Count of All Effluent Violations	0.59 (2)	0 - 73	10,585 (13%)
5	Count of Values that Exceeded 40% or 20% of Limit Value for Category 1 or 2 Parameters Respectively	0.4 (1.62)	0 - 73	10,585 (13%)
	Predicted Aggregated Time Series Features From the Target Quarter*			
6	Measurement Count Across All Parameters	17.86 (21.7)	0 - 959	9,437 (12%)
7	Unweighted Sum of Exceedance Percentages Across All Category 1 and 2 Parameters	304.71 (6,944.86)	0 - 612,493.4	9,437 (12%)
8	Weighted Sum of Exceedance Percentages Across All Category 1 and 2 Parameters	11.37 (266.45)	0 - 28,142.59	9,437 (12%)
9	Count of All Effluent Violations	0.51 (1.99)	0 - 56	9,437 (12%)
10	Count of Values that Exceeded 40% or 20% of Limit Value for Category 1 or 2 Parameters Respectively	0.32 (1.54)	0 - 56	9,437 (12%)
	Facility-Level Features			
11	The Amount of Flow (Million Gallons per Day) a permitted facility was designed to accommodate	3,173.83 (161,169.3)	0 - 19,800,000	44,881 (56%)
12	The Amount of Flow (Million Gallons per Day) that the facility actually had at the time of application	5,067.7 (439,889.7)	0 - 51,000,000	57,604 (72%)

Categorical Features

	Feature	Category Count	Mode (%)	Missing Count (%)
	Permit-Level Features			
13	Facility Type (POTW, Non-POTW, or Federal Entity)	4	Non-POTW (66%)	1,718 (2%)
14	Individual or General Permit	3	Individual (67%)	1,716 (2%)
15	Major or Minor Permit	3	Minor (83%)	1,716 (2%)
16	Wastewater or Non-Wastewater Permit	2	Wastewater (90%)	0 (0%)
17	Sewage Treatment or Non-Sewage Permit	2	Non-Sewage (62%)	0 (0%)
18	Ownership Type of the Facility (e.g. Municipality)	15	Privately Owned Facility (30%)	11,145 (14%)
19	Category of Water Body that the Permit Discharges to	2	Missing (Non-303(D)-Listed) (60%)	48,341 (60%)
	Historical Statuses			
20	The Official Historical Compliance Status One Quarter Before the Target Quarter	14	Missing (Automatic Compliant) (32%)	25,290 (32%)
21-27	The Official Historical Statuses Two to Eight Quarters Before the Target Quarter**	14	-	-

Table 4: Summary Statistics of All 27 Prediction Features in the Random Forest Model. *As the target quarter in the test set is FY2020-Q1, the previous quarter refers to FY2019-Q4. In the training dataset (target quarter FY2019-Q4), the previous quarter refers to FY2019-Q3. **The remaining historical status features share similar summary statistics and are thus omitted here.


Machine Learning Flow Chart
Raw Data
EPA Integrated Compliance Information System (ICIS) - NPDES
ARIMA Forecasts
ARIMA Forecasts
Final Merge
Final Dataset
Machine Learning Models
Random Forest Classification Model
Random Forest Regression Model

Discharge Monitoring Report (DMR) Data

This table contains DMRs which are periodic self-reports submitted by facilities to state levelenvironmental agencies (and subsequently to the U.S. EPA) withinformation about their compliance with permitted effluent limits.

NPDES Permit ID
Pollutant Parameter
Monitoring Period Date
Limit Value
DMR Value
Exceedance Percentage

Facility Data

This table contains metadata information about each NPDES Facility.

NPDES Permit ID
Location (e.g. Lat/Long, Census Block Group, etc.)
Permit Type (e.g. Wastewater vs Stormwater, Individual vs General)
EPA Region
Total Design Flow Number: The flow that a permitted facility was designed to accommodate, expressed as millions of gallons per day (MGD).

Historical Quarterly Compliance Status Records

This table contains historical compliance records (e.g. whether the facility was in SNC) on a quarterly basis for NPDES Permits

NPDES Permit ID
Quarter
Significant Non-compliance Status

A simplified view of the underlying data. With the DMR Data, we observe for each facility and parameter, self-reported discharge values in each monitoring period and can compare these self-reported values to the permitted limits.

We then merge the data with facility-level characteristics and historical quarterly non-compliance records to assemble the training data for the ML Model.

NPDES Permit ID	Pollutant Parameter	ARIMA Predicted Discharge Value
CA0001	Nitrogen	25 mg/L
CA0001	Phosphorous	10 mg/L
...	...	...

ARIMA Forecasted Results

Forecasted discharge amount for each NPDES Permit and Pollutant Parameter for 2020 Q1

NPDES Permit ID
Pollutant Parameter
ARIMA Forecasted Discharge Amount
Limit Value

Using ARIMA, we forecast the discharge amount for each NPDES Permit and Pollutant Parameter for 2020 Q1. We use these forecasted values as features in the final Random Forest machine learning model.

NPDES Permit ID	Total Number of Predicted Violations, based on ARIMA (both Chronic and TRC Violations)	Historical Quarterly Non-Compliance Status (Vector of Indicators)	Facility Level Characteristics
CA0001	4	1	Individual Permit, Wastewater, EPA Region 10, etc.
CA0002	2	0	General Permit, Stormwater, EPA Region 6, etc.
...	...	...	

Fully Merged Dataset

The final data set merges ARIMA Forecasts, Facility Level metadata, and historical noncompliance status at the permit level

NPDES Permit ID
Total Number of Predicted Exceedances (based on ARIMA Results)
Historical Quarterly Non-compliance statuses
Facility Level Metadata (e.g. Permit Type, Region, Total Design Flow Number, etc.
Predicted Discharge Value

The final dataset, after merging all of the relevant datasets, is uniquely identified at the facility level. ARIMA results at the permit and pollutant parameter level are aggregated to the permit level by taking the total number of predicted exceedances

We train both Random Forest models using 2019 Q4 data and test on 2020 Q1 data. Both the Regression and Classification models use the same feature space. The only difference is the outcome variable.

Classification Model: The output is a binary indicator for whether the facility triggered Effluent SNC.

Regression Model: The output is a weighted sum of the total exceedance percentage across Group 1 and 2 pollutant parameters for the facility.

Figure 7: The first row provides a simplified view of data drawn from the ICIS-NPDES database to construct input features in the model. The second row represents the time series model. The third row outlines the final dataset used for prediction at the permit level for both the Classification and Regression models, merging features constructed by different data sources. Finally, the fourth row feeds the prediction dataset into a Classification and a Regression model with different outcome variables.

C.2 Regression

Using the weighted sums of exceedance percentages in FY2020-Q1 as the continuous outcome variable, the RF regression model produces a root mean square error of 240, compared with a mean imputation baseline of 288. As shown in Figure 5, the regression results were driven by a small number of outliers in the last bin. (Even that error, however, still places facilities disproportionately into SNC territory.) Performance is more reasonable trimming the top 3% of outliers, resulting in an RMSE of 117.

We again assess feature importance, and ARIMA forecasts have the highest predictive power, suggesting that the ARIMA models provide some of the most useful information in predicting effluent amounts.

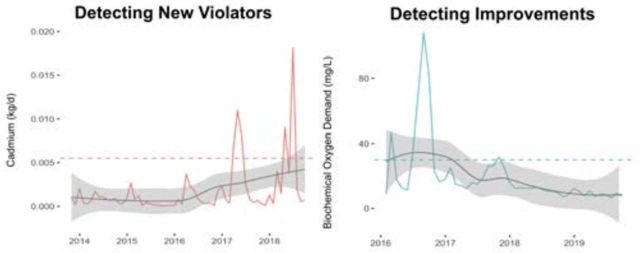


Figure 8: ARIMA-based feature engineering of parameters at NPDES facilities. Dashed line represents permit limit. Solid lines represent pollutant values over time and bands present ARIMA-based pointwise confidence intervals.

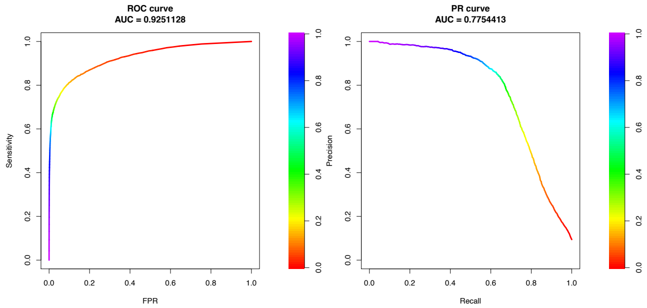


Figure 9: Left Panel: Receiver Operating Characteristics Curve (AUC = 0.93) shows that the classification model can distinguish between a true negative and a true positive fairly well. Right Panel: Precision and Recall Curve (AUC = 0.78) shows that the classification model performs reasonably well in the face of class imbalance.

Figure 10 shows that the ARIMA predictions in the permit-level center around 0. Similar to the classification model, historical statuses and time series features have higher predictive power than time-invariant permit characteristics.

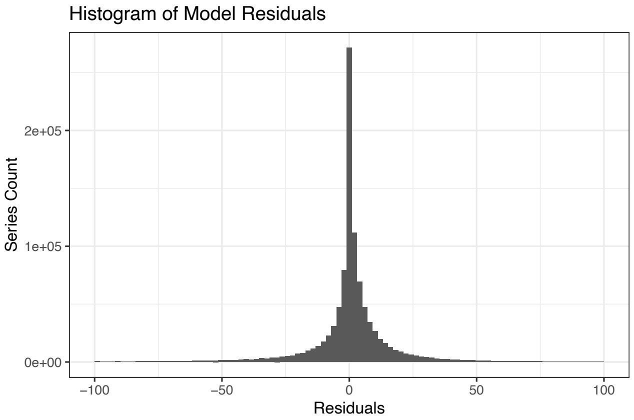


Figure 10: Residuals of the regression model center at 0.