Production and Operations Management

Efficient Detection of Environmental Violators: A Big Data Approach

Xiangyu Chang*
Department of Information Management and E-Business, School of Management, Xi'an Jiaotong University, xiangyuchang@xjtu.edu.cn

Yinghui Huang
Department of Industrial Engineering and Decision Analytics
School of Engineering, The Hong Kong University of Science and Technology

Mei Li*
Division of Marketing and Supply Chain Management, Price College of Business, University of Oklahoma, mei.li@ou.edu

Xin Bo
Appraisal Center for Environment and Engineering, Ministry of Ecology and Environment, Beijing, China

Subdoha Kumar
Fox School of Business, Temple University

*Corresponding Authors

The detection of environmental violators is critical to the long-term adoption of sustainability in supply chain management. However, there exist manufacturing facilities that report false environmental monitoring data, thereby seriously hampering governments efforts to identify true offenders and to properly intervene. We integrate waste gas data from the worlds largest Continuous Emission Monitoring System (CEMS) with a publicly available Violation and Punishment Dataset to build prediction models for the identification of environmental violators. We utilize and create innovative machine learning approaches to overcome analytical challenges associated with empirical data. First, we use a feature engineering approach to generate features from the raw, and possibly fraudulent, reporting data. This overcomes the challenges associated with low fidelity, irregularity, and the presence of extreme values in the raw dataset. Second, while building prediction models, we develop new approaches to positive and unlabeled learning to overcome the challenges posed by sparsity and mislabeled data. Our prediction model achieves satisfactory results in a related field test. Our study develops new techniques for big data analytics, which greatly improve the efficiency and effectiveness in detection of environmental violators and enhance operational outcomes of environmental protection agencies. This research is a joint effort between academia and practitioners, as evidenced by the participation of the Ministry of Ecology and Environment of Peoples Republic of China. The Ministry kindly granted us direct data access, as well as opportunities to interview Subject Matter Experts at the Ministry, which led to research insights incorporated in this manuscript. Our research findings have global implications, as CEMS devices are universally adopted to monitor waste gas emissions.

Key words: Big data analytics, positive and unlabeled learning, sustainability, violator detection
History: Received: July 2019; accepted: August 2020 by Qi Feng after one revision.

1 Introduction

Environmental violation refers to firms operational activities that violate environmental law or regulation (Karpoff et al. 2005). Some examples include the improper release of waste gas into the air and the improper treatment of hazardous water. Environmental violations are severe, global, and widespread. In the United States, the Environmental Protection Agency (EPA) detected a large number of violations by businesses in 2017, which led to over $2.9 billion in criminal fines and restitution (EPA 2019). Similarly, there were an increasing number of violations from 2016 to 2018 in Europe (ECR 2019). Environmental violation is more problematic in developing countries (Da Silva et al. 2017, Roque et al. 2018).

China, as the worlds factory (Plambeck et al. 2012), suffers severely from environmental problems (Chen 2018), which led to the mandatory installation of the worlds largest Continuous Emission Monitoring System (CEMS) by manufacturing facilities. Despite governmental efforts, inspections by the Chinese government have revealed that some manufacturers have systematically circumvented the monitoring devices and reported false pollutant data, rendering the monitoring mechanism useless. This falsification of data poses an additional challenge in identifying environmental violators. There is a dire need to find an efficient and effective approach that can work with low fidelity in reported data and systematically detect environmental violators.

Although this is an important operational problem, it has not been sufficiently researched in the operations and supply chain management (O&SCM) literature. Recently, researchers have advocated the use of data-driven analytics to solve challenging O&SCM issues (Choi et al. 2018, Cui et al. 2018, Guha and Kumar 2018, Sanders and Ganeshan 2018, Shmueli and Yahav 2018). In the sustainability research space, various scholars point out the need to use big data analytics in the investigation of complex environmental issues (Singhal et al. 2018, Wang et al. 2016). In this paper, we respond to these calls by conducting research on the detection of environmental violators, based on big CEMS data obtained from China. Because CEMS devices are universally adopted to monitor waste gas emissions, our research findings have global implications. In addition, we develop and apply innovative tools to overcome challenges associated with empirical datasets. As such, we also contribute to research in data analytics.

1.1 Research Motivation

The problem with quality of reported data has been a recurrent one. As early as the 1990s, the US federal authorities found evidence of tampering with the results of environmental tests and ordered thousands of environmental safety tests conducted between 1994 and 1997 to be repeated (Oppel Jr 2000). What is disconcerting is that in recent years, the number of exposed cases of inaccuracy in

reported environmental data seems to be increasing. The emissions test scandal stormed through the automobile manufacturing industry, and auto giants such as Volkswagen (Rhodes 2016, Siano et al. 2017) and Nissan (Hermanns et al. 2018, Gale 2018) have been found guilty of altering auto emissions test data in order to bypass environmental regulations. Similarly, Ford is facing a criminal probe into emissions testing fraud (AN 2019). The inaccuracy of mandatorily reported environmental data is not isolated to one industry. In April 2019, researchers from Environment Canada conducted an independent study and found that a number of major oilsands operations in northern Alberta seem to be emitting significantly more carbon pollution than companies have been reporting (Dubinsky 2019, p. 1).

The causes of low quality in reported data vary. They certainly include intentional falsification of the data for profit-driven reasons, as was the case for both Volkswagen and Nissan. Research shows that when there is a conflict between the objective of a public policy and the internal efficiency requirements of a firm, the firm may decouple (Meyer and Rowan 1977) its ceremonial conformity to a public policy from its actual implementation of such a policy (Boxenbaum and Jonsson 2017). In our research context, the mandatory reporting of environmental data may not be aligned with the interests of some manufacturing facilities and decoupling may occur in the form of falsifying reported data. In general, the falsification of data reduces firms necessary investments in environmental compliance and reduces the operational costs for the offending firms. Besides intentional abuse in reporting data, low quality may be caused by unintentional negligence with regard to data, as was the case with oilsands operations in Canada where firms may have reported inaccurate environmental monitoring data due to incorrect procedures (Dubinsky 2019).

Regardless of the causes, low quality of reported environmental monitoring data can pose severe threats to the public. Environmental agencies rely on environmental testing data to monitor and manage manufacturing facilities, and to intervene when necessary. The reporting of low-quality data misguides governmental decisions and distorts the efficiency and effectiveness of governmental operations. Furthermore, the damaging effect goes beyond the distortion of governmental operations. Low quality in reported data conceals the true identify of environmental violators and harms the triple bottom line of sustainable operations (Hussain et al. 2018). Environmentally, it endangers public health and the global ecosystems (Lee and Xiao 2020, Mendelssohn et al. 2012). Socially, environmental violations can be one of the major causes of tension between businesses and local residents (Calvano 2008). Economically, environmental violations not only weaken a violating firms financial health and stability by incurring a significant amount of fines and penalties (Xu et al. 2012, Zou et al. 2015), they also impede fair competition for non-violating firms (Delmas and Keller 2005), which incur higher costs in order to abide by environmental regulations. It is

imperative that these violators be detected and caught. This is critical for the long-term adoption of environment improvement initiatives (Elkington 1998).

Extant research, though it has recognized the damaging effects associated with environmental violations (Delmas and Keller 2005, Lee and Xiao 2020, Mendelssohn et al. 2012, Zou et al. 2015), has largely neglected the important topic of mechanisms for detection of environmental violators. A very limited number of studies on environmental violations take a legal and/or public administration perspective, and these are restricted to the creation of laws and regulations to guide firm behaviors (Karpoff et al. 2005). There are very few (if any) studies that investigate the actual implementation of these laws or regulations, including finding powerful detection mechanisms to discover violations in order to efficiently enforce such laws/regulations. Even more rare is research that can work with a massive amount of raw environmental reporting data, detect anomalies in reporting, and discover hidden or concealed environmental violators. As a result, such detection relies heavily on manual inspections to catch cheaters who intentionally manipulate the environmental monitoring devices and/or alter reporting data. Yet sporadic inspections are expensive and have a low detection rate. The environmental protection agencies can benefit from a more targeted and systematic mechanism to efficiently catch violators. The focus of this study is to develop such a mechanism. In doing so, we must overcome considerable challenges associated with the raw empirical dataset by creating and applying new machine learning methods to enable analysis of the dataset, and to eventually make accurate predictions based on it.

1.2 Research Questions and Contributions

Researchers in data-driven analytics have recognized both the opportunities and the challenges brought by big data and business analytics (Choi et al. 2018, Corbett 2018, Lee and Xiao 2020, Sanders and Ganeshan 2018). In the context of environmental research, scholars call for the use of big data to support effective and efficient decision-making on sustainability issues (Wang et al. 2016). Our research answers these calls and pioneers a data-driven analytics in the context of detection of environmental violators. Below we present our key research questions and contributions.

We begin by asking our first set of research questions (RQs). RQ (1): *Given that pollutant reduction often works against firms economic interests, how do we know if the environmental-monitoring data reported by manufacturing firms are of high quality? Can we create a tool that can be used to measure the quality of reported data?* In our research context, we use the term reporting data quality to denote the level of fidelity (or lack of fidelity) in reported environmental-monitoring data. Although we recognize that in the generic sense, data quality could mean many things (Dey and Kumar 2010, 2013), we adopt a narrow definition here to fit our research context. We develop a reporting data quality assessment framework that is made up of a set of reporting data quality

indicators, collectively called a reporting data quality index. This index can work with the massive, often messy (Ozdemir and Susarla 2018), and unreliable nature of raw environmental pollutant monitoring data, capture its key features, and measure its quality.

The unique contribution of this reporting data quality index goes beyond the detection of inaccuracy caused by noise in big data. It is capable of detecting low fidelity caused by deliberate deceptions (Rubin and Lukoianova 2013). This is a major departure from previous research. Specifically, the conventional research typically operates directly on raw data. This is feasible because it assumes the underlying trustworthiness of the raw data. Even though incomplete, incorrect, inaccurate, or irrelevant data may exist (Deepa and Chezian 2014), it does not undermine the fundamental trustworthiness of the dataset. In this regard, researchers are mostly concerned with deriving efficient algorithms to prewash raw data (Dey and Kumar 2013) before further decision analysis, such as preprocessing and filtering to avoid obvious irrelevant information (Yin and Kaynak 2015, p. 145), and designing algorithms to compute missing data or replacing inaccurate data (Kumar and Chadrasekaran 2011). None of these techniques would work without the underlying assumption of overall trustworthiness of the dataset. However, this assumption does not work in our research context, or in any other research context where attempts have been made to intentionally distort the reported data, such as in the aforementioned cases of Volkswagen or Nissan. The question then is how *not to* trust the raw data, yet be able to detect from it *reliable clues* that inform us on the quality of reported data.

We accomplish this tall order by utilizing a feature engineering technique. Feature engineering allows us to make use of domain knowledge to extract features from the raw data for better business decision-making (Brownlee 2014, Kumar et al. 2018, 2019). In essence, the feature engineering technique helps us to extract features, i.e., metadata from CEMS reporting, to generate a reporting data quality index. This data quality index measures the fidelity of the reported data and provides important clues to possible falsification behavior by dishonest environmental violators. Thus, it affords us the ability to detect data quality problems associated with deliberate deceptions. Our reporting data quality index is particularly useful in mandatory reporting situations where there may be conflict of interest between the public administrative offices and private reporting firms (Marquis and Qian 2014). While we do not claim a theoretical contribution to decoupling theory, our research findings provide an effective tool to detect the existence of decoupling in an environmental reporting context.

Our second research question builds on the reporting data quality index and is predictive in nature. RQ (2): *How can we make use of the reporting data quality index and make predictions on environmental violators? Which predictive model works best?* To answer this question, we integrate the reporting data quality index derived from CEMS data with a publicly available Violation and

Punishment Dataset (VPD). VPD contains a list of manufacturing facilities that, through manual inspection, have been found to have violated environmental regulations. Thus, VPD informs us on the confirmed status of violations. Linking VPD with the data quality index, we can build prediction models to catch cheaters, i.e., facilities with a high probability of manipulating the reporting devices and submitting falsified emission data.

Because only a small number of cheaters have been caught by the manual inspection process, there are two challenges in building predictive models. First, the matrix denoting the status of the cheaters is sparse. Due to the high cost and the labor-intensive nature associated with manual inspections, only a small number of violators exist in VPD. Second and relatedly, there also exist a large number of unidentified true violators. In this study, we develop two new approaches to positive and unlabeled learning (i.e., PU learning) to overcome these two challenges and improve the predictive accuracy of our models.

Answers to our second research question make the following contributions. First, we develop an efficient and accurate prediction model that can be applied to the detection of environmental violators from CEMS data. Given that CEMS is a popular device adopted by environmental protection agencies globally for the monitoring of pollutants (ECCC 2020, EPA 2020, Gupta 2019, Zhang and Schreifels 2011) with market growth projected to reach $4.44 billion by 2025 (Danigelis 2018), our findings would benefit current as well as future environmental protection efforts worldwide. As such, we contribute to sustainable operations research in O&SCM by addressing the implementation side of sustainability enforcement. Second, methodology-wise, we develop new algorithms that contribute to the analysis of positive and unlabeled data (Li and Liu 2003). These algorithms are extremely helpful in detecting patterns in an empirical dataset that contains a large amount of unlabeled data. Third, our research also has significant managerial implications. Our prediction of offenders fundamentally improves the monitoring and detection mechanism for environmental protection agencies. The new detection process can potentially lead to substantial cost reduction and improved detection efficiency.

2 Literature Review

Here, we briefly discuss the literature in the following streams: (a) sustainable operations, (b) challenges in deriving business insights from big data, and (c) descriptive, predictive, and prescriptive data analytics. The goal of this section is to not only review existing literature, but also to highlight our contributions with respect to existing literature. At the end of the section, we contrast explanatory models and data-driven predictive models, and explicitly ground our research in the latter stream.

2.1 Existing Research on Sustainable Operations

Online Appendix 1 summarizes existing research on sustainability and environmental violations. As Online Appendix 1 shows, although there are an increasing number of publications on environmental sustainability in the O&SCM field (Bernard et al. 2018, Goebel et al. 2018, Hussain et al. 2018, Porteous et al. 2015, Xia et al. 2018), the key focus has been on answering the question of how to promote sustainability throughout the supply chain, and limited research has also addressed the possible negative consequences of non-compliance (Xu et al. 2012, 2016, Zou et al. 2015). With the damaging effects of non-compliance in mind, research has also shed light on the mechanisms to contain such behaviors (Gray and Shimshack 2011). Besides empirical research summarized in Online Appendix 1, there are also important review papers that identify new research opportunities (Lee and Tang 2018, Tang 2018), as well as conceptual papers that explain novel concepts in sustainability (Arenas and Rodrigo 2016, Joglekar et al. 2016, Murray et al. 2017, Nonet et al. 2016, Sodhi 2015).

A careful examination of the existing literature revealed a serious gap. Although past studies have gained a preliminary understanding of what drives environmental compliance and the importance of environmental monitoring and enforcement (Gray and Shimshack 2011), there is very little research on how to effectively operationalize environmental monitoring and enforcement efforts. Specific to our research context, we are not aware of any research that focuses on the effective detection of environmental violators. As a result, environmental violation detection operates largely on an inefficient inspection basis, which results in a low detection ratio and high monitoring costs. We believe that public policy enforcement can benefit from tools and principles in operations management research to improve the efficiency of violation detection. Specific to our research, we borrow tools from big data analytics to improve efficiency in violation detection.

2.2 Challenges in Deriving Business Insights from Big Data

Big data is commonly characterized by four Vs: volume, variety, velocity (Choi et al. 2018, Mishra et al. 2018, Rozados and Tjahjono 2014), and a somewhat recent addition, veracity (Guha and Kumar 2018). Among the four characteristics, the first V, volume, is arguably the most defining feature of big data (Labrinidis and Jagadish 2012).

Volume refers to the large number of observations available for analysis. While volume of data presents the potential for deriving useful business insights, the transformation is not automatic (Jagadish et al. 2014). As raw data are accumulating at unprecedented rates, they need to be organized, prepared, and analyzed before business intelligence can be extracted from the data. In this regard, researchers have observed a growing gap between data and users (Kepner et al. 2014)

and call for innovative ways to address this challenge. In our research, we take a massive amount of raw reporting data on environmental monitoring and apply a feature engineering technique to derive a set of key features of such data. These key features make up the data quality index. This index measures the fidelity of reported data and is instrumental in data consumers evaluation of the credibility of reporting firms. More importantly, this reporting data quality index serves as a good predictor for the detection of environmental violators. Thus, our research helps transform a massive amount of raw data into meaningful operations management decision insights.

2.3 Descriptive, Predictive, and Prescriptive Data Analytics

Due to the newness of big data analytics in O&SCM, the definitions for some terms are still evolving. Yet, a common way to describe and distinguish big data analytics is to divide it into descriptive, predictive, and prescriptive categories. Deka (2016) provides detailed descriptions of the three terms.

Descriptive analytics is past-oriented. The objective is to analyze historical data in order to identify patterns. For example, Foster et al. (2018) perform cluster analysis to identify patterns that impact emergency room physicians' performance. Li et al. (2016) use inverse covariance estimation technique to detect operational patterns associated with high performing manufacturing firms. Predictive analytics also analyzes past data, with the intention to predict future outcomes. For example, Cui et al. (2018) mine social media data to make predictions on future sales. Similarly, Boone et al. (2018) use Google trends information to improve sales forecasts. Finally, prescriptive analytics provides recommendation for future actions. IBM treats prescriptive analytics as the final phase of big data analytics as it evaluates and determines new ways to operate (Deka 2016).

Online Appendices 2 and 3 summarize research that utilizes big data analytics to solve O&SCM challenges. The existing O&SCM research on descriptive, predictive, and prescription analytics typically focuses on one of the elements (Cui et al. 2018, Lau et al. 2018, Li et al. 2020). Very few studies have touched on all three aspects. One exception is Swaminathan (2018) who describes, conceptually, how all three types can be used in humanitarian operations. In this paper, we build on Swaminathan (2018) and illustrate how we operate descriptive, predictive, and prescriptive analytics on empirical datasets to solve an important environmental enforcement problem. Specifically, we first utilize descriptive analysis to describe the features of the reporting data. These features are critical to the detection of environmental violators, as they serve as the inputs to our prediction models. Second, we build prediction models utilizing recent machine learning techniques. We also develop new PU learning models to overcome challenges in our empirical datasets and produce the best-fitting one for the most accurate and efficient detection of environmental violators. Lastly,

as a part of the validation efforts of our prediction model, we prescribe a new data-driven business process to catch environmental violators. This new process achieved satisfactory results in a related field test. Although our research centers on the prediction model, we have demonstrated the complete range of descriptive, predictive, and prescriptive elements of big data analytics. As far as we know, we are the first to do so in research on sustainable operations management.

2.4 Data-Driven Prediction Models vs. Explanatory Models

We take note here that unlike the traditional research mode with a primary focus on building explanatory models that identify casual relationships among model constructs, a distinctive feature of data-driven predictive analytics is its focus on prediction (Boone et al. 2018, Cui et al. 2018) and its application in decision-making. In this regard, the seminal work of Shmueli and Koppius (2011) provides a detailed explanation of the differences between explanation models and prediction models in terms of analysis goal, variables of interest, model building optimized function, model building constraints, and model evaluation. Similarly, Kitchin (2014) echoes the view of Shmueli and Koppius (2011). In his influential piece on data-driven analytics, Kitchin (2014) treats big data analytics as disruptive innovations, and therefore calls for a paradigm shift on how research is done with big data. In line with Boone et al. (2018), Cui et al. (2018), Shmueli and Koppius (2011), and various other papers published in the two special issues of Production and Operations Management on Big Data Analytics, our research focus is prediction and its impacts on decision-making. We do not intend to build explanatory models or to claim casual relationships between predictors and outcome variables. Thus, we root our study in the general framework of data-driven analytics. Below we explain our research context.

3 Research Background and Research Data

3.1 Research Background

China has experienced phenomenal development triggered by the economic reform in the late 1970s (Chow 1993, Holz 2008, Singhal and Singhal 2019). As an emerging economy, China initially formulated a policy that prioritized economic development, which resulted in rapid deterioration of the environment (Fang et al. 2009), including massive air and water pollution (Ebenstein 2012, Ji et al. 2014). In recent years, the central government has made strategic decisions to reduce pollution and improve the environment (Li et al. 2015).

One of the key resolvents to ongoing environmental deterioration in China is the mandatory installation of CEMS in polluting facilities, following the lead of the United States (Pan et al. 2005, Zhang and Schreifels 2011). CEMS is employed worldwide (ECCC 2020, EPA 2020, Gupta 2019) to directly and continuously monitor, record, and report the emissions measurement and

operating parameters for required pollutants, such as concentration of emissions and discharge flow. By the end of 2016, the total number of CEMS installed in China exceeded 29,000. The installation of CEMS in China covers various industries such as thermal power, iron and steel, metallurgical, aluminum, chemical, waste-water treatment, and paper-making industries (Tang et al. 2019). The Ministry of Ecology and Environment of the Peoples Republic of China (MEEC) mandates pollutant-producing facilities to continuously report pollutant monitoring data (Karplus et al. 2018). Under this monitoring system, if the average hourly concentration of a pollutant exceeds a preset limit five times, an inquiry will be made by the agency. Without a reasonable explanation, a public investigation will be opened and possibly a penalty will be issued. Over the years, this CEMS dataset has accumulated a tremendous amount of data.

Even with the worlds largest and most comprehensive monitoring system, there exist many violations. There are firms that circumvent the monitoring devices and release more pollutants than the allowance. Some examples include destroying monitoring devices, modifying monitoring data, and diluting pollutant intake. This type of violation is so severe and prevalent that the Chinese government had to set up an additional manual inspection mechanism to catch the cheaters. Each year, employees from MEEC are pulled from their regular posts to conduct inspections at the provincial level. This prevents them from carrying out their normal job duties. At the same time, it incurs a high travel-related cost. Due to the high costs and resource drain associated with manual inspection, it is largely carried out on a random basis with occasional inspections based on tips received from anonymous phone calls.

Overall, the current inspection mechanism is expensive and has a low detection rate. However, we note that over the years, these inspections have resulted in the accumulation of a list of companies that seemingly report normal, passing-grade pollutant data, yet are caught violating the environmental regulations. This dataset of environmental violators is publicly available. Yet we are not aware of any research that integrates this dataset with the big CEMS data and explores a possible connection between these two sets of data. Our research makes the first attempt to take full advantage of the integration of these two datasets and build prediction models to identify violators. Below we describe these two datasets.

3.2 Data Description

As discussed earlier, we utilized two datasets: The first is the CEMS dataset and the second is the VPD. Briefly, the first dataset includes all the information on daily gas pollutants emission activity, reported by manufacturing facilities. The second dataset includes the list of firms that have been detected as environmental violators through inspections and subsequently penalized by MEEC.

Table 1 Summary Statistics for Factories and Pollutants Included in CEMS Dataset

Measure	Pollutants Smoke Dust	SO_2	NO_x	COD	Ammonia-nitrogen
Number of discharge outlets	18,257	18,257	18,257	11,770	11,770
Number of emission records	159,977,477	159,977,477	159,977,477	122,854,626	122,854,626
Average concentration	77.63	21.05	118.87	46.77	3.13
Median concentration	24.29	10.6	70.09	27.59	0.98
Standard Deviation	143.90	59.54	141.80	69.31	10.60

Table 2 A Simplified Example of CEMS Record

Factory ID	Outlets ID	Nature of Factory	Pollutant	Time	Concentration	Flow
10000002	003	Small scale, Private enterprise, Non-key-control	NO_x	05/12/2016, 11:00	22.80	43.06
10000001	002	Large II scale, State-owned factory, Province-control	SO_2	05/12/2016, 11:00	18.06	26.17

3.2.1 CEMS Dataset The CEMS dataset is provided by MEEC and obtained directly from the national CEMS monitoring data platform. The dataset includes CEMS hourly pollution emission data for all industries in 30 province-level regions of mainland China from January 1, 2016 to June 30, 2017. In total, the CEMS dataset includes monitoring data from 7,643 factories. There are 725,641,683 emission monitoring records for five major pollutants: smoke dust, SO_2, NO_x, COD, and Ammonia-nitrogen. The summary statistics of CEMS data are listed in Table 1.

The CEMS records the maximum, minimum, and average value of emissions concentration (mg/m^3 for gaseous pollutants) hourly, and then automatically revises updated concentration and flow. In addition, the CEMS also measures necessary operating parameters, such as the oxygen content for boilers, which provide valuable information on the state of the equipment. It is worth mentioning that the combination of pollutants being constantly monitored by CEMS varies by industry, but it covers all major pollutants for each industry. The thermal power industry, for example, is required to report the emissions data for smoke dust, SO_2, and NO_x, whereas municipal sewage plants report data for emissions of COD and Ammonia-nitrogen instead. The CEMS assigns a unique ID number for each discharge outlet. Thus, we are able to match factory information with its corresponding hourly monitoring records through the ID number. A snapshot of the CEMS dataset is provided in Table 2. Note that in the actual dataset, billions of records are generated each year, including both gas and water pollutants. Our study mainly utilizes gas pollutant data.

The nature of the factory column tracks three types of information including scale of the factory (from small to super large), registration type, and emission control level (state-controlled, province-controlled, city-controlled, and non-key polluting facilities). Since the reporting of gas pollutant data is mandatory in China, missing data are at a minimum. We only encountered 8 missing records in registration type, 7 missing records in the scale of factories, and 64 missing records in control level. We subsequently removed these missing records.

3.2.2 VPD To build prediction models, we integrate the CEMS dataset with VPD. The manual inspection mechanism has helped MEEC to detect a number of violators and subsequently penalize them. We assemble VPD from the publicly available online data platform hosted by MEEC. All environmental-related administrative penalties are released on the platform quarterly (http://www.mee.gov.cn/gkml/?ClassInfoId=119), allowing us to link the CEMS data with factories actual status of compliance with environmental policies. VPD contains 372 factories with administrative penalties issued during 2016-2017, which includes 118 factories with waste gas pollution violations.

4 Problem Formulation and Empirical Challenges

Our objective is to create a predictive model for the efficient detection of environmental violators. This predictive model utilizes each firms CEMS reporting data and matches them against violation status as stored in VPD. On the surface, this problem can be formulated as a classic binary classification problem as denoted below.

Suppose that there are n factories in the CEMS database, and the i-th factory has been recorded n_i times as $\mathbf{r}_i = (r_{i1}, r_{i2}, \ldots, r_{in_i})$. According to the VPD, we assume that $l_{it_i} = 1$ if the i-th factory is punished at timestamp t_i, where $i \in \{1, \ldots, n\}$ and $t_i \in \{1, \ldots, n_i\}$. At first glance, one can assume this is a supervised classification problem by considering $\mathbf{r}_i$ as the feature of the i-th factory and l_{it_i} as the corresponding label. However, this classic machine learning approach faces the following challenges posed by the special characteristics of the CEMS dataset.

Irregularity. First, different pollutant-producing factories began to operate the CEMS devices at different time intervals. This difference in turn leads to a different number of records produced by different factories. This irregularity presents a challenge for classical models such as Logistics Regression, AdaBoost, and Random Forest, which all require a regular table as their input (Friedman et al. 2001). Therefore, it would be a violation of model assumption to directly apply classic supervised machine learning models to the raw data.

Outlier. An outlier is an observation point that is distant from other observations, and an outlier can prevent researchers from obtaining robust results (Huber 2004). Outliers exist in the CEMS dataset. They can be caused by measurement error, unstable operation, or abnormal production process. The frequent existence of outliers also poses a problem for the direct application of classic supervised learning method, i.e., the robustness of the findings can be jeopardized (Lecué et al. 2020).

Sparsity. One critical challenge in applying the classic supervised machine learning model is that the punishment label l_{it_i} is very sparse. Although we have millions of records on the predictors obtained from the CEMS dataset, we only have a very limited number of records of outcome

variables, i.e., $l_{it_i} = 1$ for the factories that were caught by MEEC. This creates severe sparsity in the matrices.

Missing Label. Related to the problem above, the sparsity of the data matrix also means that there may exist many violators who report falsified environmental monitoring data yet have not been caught, due to the resource constraints associated with the manual random inspection process. If we treat the violation status of a firm purely based on information from VPD (i.e., if a facility exists in VPD, set violation status $= 1$, and -1 otherwise), it is very likely we would have mislabeled many facilities as -1 although they have in fact violated environmental regulations (but have not been caught by the inspection). In other words, if a facility is caught by MEEC, it is a violator. However, if a facility is not caught by MEEC, it does not mean it is a non-violator. Thus, this leads to a number of missing labels in the raw data. If we were to label non-punished facilities with the label $l_{it_i} = -1$, it may result in large biases in decision-making due to the incorrect labeling.

Lack of Fidelity in Reporting Data. As we allude to earlier, a major problem in the raw CEMS reporting data is the low fidelity caused by intentional manipulation of the reporting devices. This problem undermines the underlying trustworthiness of the raw data. Therefore, it is not advisable to operate directly on the raw data. Rather, we need to find some reliable clues, not from the raw data itself, but from the features derived from the raw data, to detect violators.

To overcome these challenges, we endeavor to leverage two methods. First, we use feature engineering to abstract features of the reported raw data and use these features to construct a new CEMS data quality assessment framework. This is consistent with Cormack et al. (2007) and Garla and Brandt (2012), who show that utilizing features instead of raw data results in better predictive power in their respective research contexts. Our data quality assessment framework incorporates six data quality features that can be used to evaluate the fidelity of the raw data reported. These six data features can overcome the low fidelity, irregularity, and outlier issues for CEMS data. Second, we propose two new PU learning methods to conquer the sparsity and missing label issues associated with raw CEMS data. The details are provided in the following two sections.

5 CEMS Data Quality Assessment Framework

In this section, we propose a new data quality assessment framework that includes a number of features derived from CEMS data. We derive these features from two sources. One is through surveying existing data quality literature. The other is through domain knowledge gained from qualitative interviews with subject matter experts (SMEs). Below we describe the construction of this data quality assessment framework.

5.1 Data Quality Literature and Qualitative Interviews

Existing literature has long recognized that data quality is a multi-dimensional concept (Ballou and Pazer 1985, Dey and Kumar 2010, Wang and Strong 1996). Many authors have provided definitions of the dimensions of data quality from different perspectives. Strong et al. (1997) define data quality as the fitness for use by data consumers. Ballou and Pazer (1985) suggest that accuracy, completeness, consistency, and timeliness are the dimensions of data quality. Wang and Strong (1996) provide an analysis from the perspective of those who use the data, and group various attributes of data quality into intrinsic, contextual, representational, and accessibility data quality classes. A detailed review is provided by Keller et al. (2017). To fit our research context, we take in the elements discussed in existing literature such as completeness, consistency (Ballou and Pazer 1985), and fitness for use by data consumers (Strong et al. 1997).

We also supplement the literature review findings with domain knowledge gained from qualitative interviews with SMEs from MEEC as well as the CEMS device manufactures. These planned interviews took place over the course of one and a half years and were scheduled quarterly. Thus, we met and interviewed SMEs at six different times throughout the project, with the first four interviews focusing on learning from SMEs and the last two interviews focusing on soliciting feedback on our proposed big data analytical approaches. At each interview, which typically lasted 2–3 hours, there were at least six experts (this was a legal requirement by the project sponsor, MEEC, to ensure the quality of the interviews). Because we wanted to obtain a comprehensive view of the issue at hand, we rotated SMEs throughout the project. Altogether we interviewed 10 unique SMEs. These experts were composed of two groups. One group included employees holding senior positions at MEEC, such as the Deputy Director of the Appraisal Center for Environment and Engineering, and the Director of Technology Support Center for Regulatory Modeling. These senior employees work out of MEEC headquarters in Beijing. (On occasion, we also invited employees holding a senior position at the provincial level.) These senior employees all have more than 10 years of work experience in the industry. This group of SMEs have a comprehensive view of the issue at a national level and shared information regarding the severity of the problem, its damages, the high costs associated with manual inspection, etc.

The second group included specialists from the CEMS device manufacturers (we selected specialists with at least 5 years of work experience), and the front-line employees for the environmental protection agencies at the provincial level and/or municipal level. Because there are many different manufacturers of CEMS devices, interviews with equipment specialists from device manufacturers informed us on conditions affecting the stability of these devices. Interviews with the front-line

employees from the environmental protection agencies provided experiential insights on some commonly used methods to manipulate the CEMS device in order to produce fraudulent reports. They also provided insights on the indications of such fraudulent activities.

We conducted the interviews with a semi-structured format (Bartlett et al. 2006). A semi-structured interview allows the researchers to control the research focus, yet at the same time allows the emergence of new ideas (Cachia and Millward 2011). The discussions were centered around the following key questions: First, which of the abnormal data in the dataset are caused by defects of the CEMS devices themselves, and which are true anomalies? Second, do these abnormal data represent fraudulent activities in reporting? And third, what are the underlying reasons causing the abnormal reporting? Information collected through these interviews proved to be instrumental in the development of the reporting data quality index. Besides the three key questions, we also allowed SMEs to expand on related topics. At the end of each interview, each attending SME signed off on the interview notes.

Our interview data suggest that at times, environmental violators do not want to be monitored by the government. When the mandatory monitoring requirements are enforced, these violators do not willingly upload high quality reporting data. For example, one of the SMEs informed us that many companies with environmental violations would deliberately cut off the power of CEMS devices when releasing toxins, and the monitoring device would temporarily fail, resulting in a data upload value of 0 or NA (not applicable). Therefore, a useful reporting data quality in our research context must factor in data quality issues caused by deliberate attempts to alter data. We combined these interview findings with data quality dimensions reported in past literature and came up with a new set of domain-knowledge-driven features derived from raw reporting data. These features can inform us on the reporting data quality suitable to our research context.

5.2 Feature Engineering and Composition of Reporting Data Quality Index

Feature engineering refers to the process of utilizing domain knowledge to extract attributes (i.e., features) from raw data, and it can be done either automatically, semi-automatically, or by manual selection (Gaber et al. 2020). Arguably a manual feature selection based on domain knowledge is preferred (Moro et al. 2014). Our research utilizes manual selection of features. We combine findings from existing research and qualitative interviews with SMEs, and develop a reporting data quality index. This index is composed of a set of six features that describe the characteristics of reported raw data. These data features can be categorized into three dimensions: Accuracy, Completeness, and Industry Particularity. The accuracy dimension captures the validation of the data, and thus is represented and measured by the abnormal ratio and zero ratio of CEMS data. The completeness dimension refers to the extent to which data are not missing when the CEMS is in operation, and

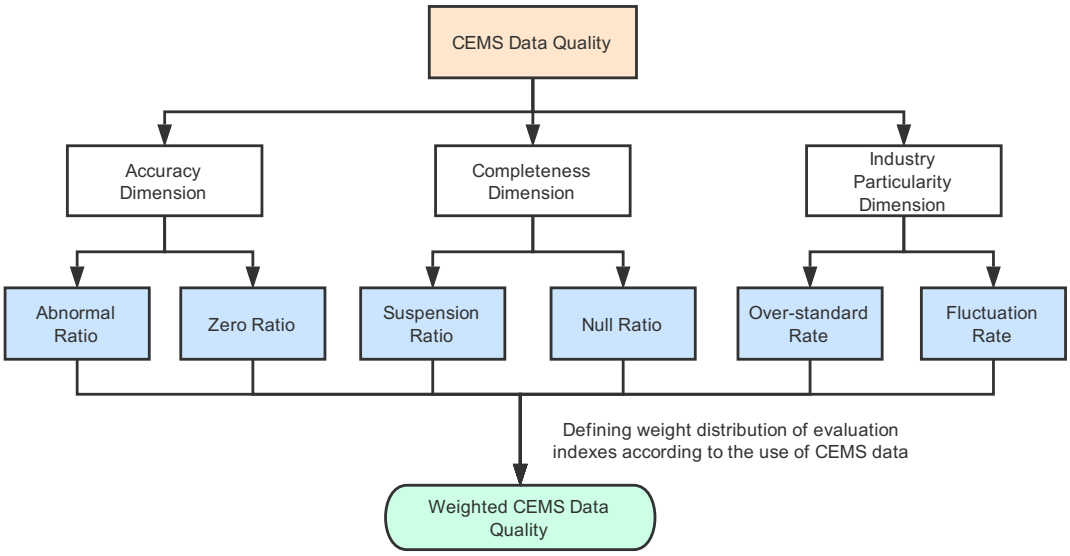

Figure 1 Reporting Data Quality Framework

is quantified by suspension ratio and null ratio. The third dimension is industry characteristics. It takes into consideration variations among industries, and measures whether the data collected are true and credible in specific industry scenarios. The framework is presented in Figure 1 and the detailed explanations for each of the six features are shown below.

5.2.1 Accuracy Dimension: Abnormal Ratio There exist negative values and abnormally large values in CEMS data. One of the key reasons is that some facilities failed to properly conduct routine calibration and reconditioning of CEMS devices. These abnormal records are not valid measurements for emissions. We use the accumulated period, which refers to the total length of negative values and abnormally large values in CEMS monitoring records, and its proportion in accumulated running time as an indication of CEMS data quality.

$$q_1 = \frac{\text{Accumulated Abnormal Period}}{\text{Accumulated Running Time}}. \tag{1}$$

Based on this formulation, we note that a large abnormal ratio signals carelessness in maintaining reporting devices and can lead to reduction in the quality of reported data.

5.2.2 Accuracy Dimension: Zero Ratio Zero values refer to scenarios in which the emission concentrations recorded by CEMS are exactly equal to zero. In reality, even if the true emission concentration is very low, the data should fluctuate around zero rather than stick to the zero position. Furthermore, one of the SMEs provided us one possible reason why continuous zero values are recorded: The manufacturing facility cuts off the power of CEMS devices when it is releasing toxins. Thus, the continuous appearance of zero values is an indicator of possible tampering with

environmental monitoring devices. We use the accumulated zero value period, which refers to the total length of zero values in CEMS monitoring records, and its proportion in accumulated running time, as an indication of CEMS data quality.

$$q_2 = \frac{\text{Accumulated Zero Value Period}}{\text{Accumulated Running Time}}. \tag{2}$$

This second indicator, q_2, is another measurement of the accuracy dimension of CEMS data, factoring potential behavioral issues of the reporting facility.

5.2.3 Completeness Dimension: Suspension Ratio We denote a suspension ratio (q_3) to reflect the stability of the device operation as follows:

$$q_3 = 1 - \frac{\text{Accumulated Running Time}}{\text{Supposed Running Time}}. \tag{3}$$

The "supposed running time" in the above equation refers to the length of time from the beginning to the end of normal operation of CEMS devices during a data quality observation period. If the suspension rate is greater than 0, we can postulate that the device was temporarily suspended for a period of time. High suspension rate implies that CEMS devices might be suffering from unstable daily operation, while a rate close to zero signals stable operations.

5.2.4 Completeness Dimension: Null Ratio There exist NA (not available) values in the CEMS data, which means that the CEMS equipment could not upload valid data. Thus, we can use the following null ratio (q_4) to reflect facilities daily operations behavior:

$$q_4 = \frac{\text{Accumulated Null Period}}{\text{Accumulated Running Time}}. \tag{4}$$

Here, "accumulated null period" refers to the total length of null value in CEMS monitoring records. A high null ratio may indicate that the factories do not follow the regular maintenance procedures and quality assurance activities for CEMS, which can be regarded as a signal of unhealthy routine operation. We cannot rule out the possibility that the CEMS device was sabotaged by the facility in order to avoid emission monitoring by the authorities.

5.2.5 Industry Particularity Dimension: Over-standard Rate The over-standard rate, which is denoted by q_5, is a ratio that measures proportion of time that a reporting device exceeds an emission standard.

$$q_5 = \frac{\text{Accumulated Out of Control Period}}{\text{Valid Time}}. \tag{5}$$

To eliminate the impact of abnormal records, we remove the period when abnormality occurred from the accumulated running time. Therefore, the valid time is defined as the accumulated running time minus the abnormal period and zero value period. The definition of out-of-control period is the number of hours that the average emission concentration exceeds corresponding emission standards. The over-standard rate gives a snapshot of reporting facilities intentions to comply with environmental policies.

5.2.6 Industry Particularity Dimension: Fluctuation Rate We noticed instances of extremely high or extremely low fluctuation of CEMS. For instance, the emission concentration reported appears to be a constant for an extended period. In this case, the variance of emission concentration may be lower than the industrial standard variance. In addition, a high fluctuation may occur when the CEMS probe suffers from poor quality and becomes oversensitive, resulting in drifts in measured values. We define the fluctuation rate as follows:

$$q_6 = \frac{|\text{Variance} - \text{Industry-standard Variance}|}{\text{Normalization Constant}}. \tag{6}$$

Here, the industrial standard variance is a self-selected reference value, and one can easily choose the average variance or median or other reasonable values for different industries.

As we discussed earlier, the lack of fidelity, irregularity, and the existence of many outliers in the raw data prevent us from operating directly on the raw data. We overcome these challenges by deriving a set of features based on domain knowledge. Specifically, transforming raw data into a uniform set of features ensures that the number of features is the same across different factories. In addition, aggregating the raw data to obtain the six data features significantly reduces the variability of raw data, which can improve the robustness of the learning models. In summary, although operating on features may lose some information in the raw data (remember that some facilities have manipulated the raw data so some reported raw data is not trustworthy), it enables us to utilize machine learning approaches to build predictive models (see Section 7). Below we describe method innovations that enable us to build such predictive models.

6 Research Method: PU Learning Framework

6.1 Why PU Learning?

Violator detection, in essence, is a binary classification problem. Our objective is to predict (classify) a facility as either a violator or not a violator. In a traditional binary classification problem, the decision function is trained by a dataset that contains both positive and negative samples, and that is fully labeled. However, in the combined datasets of CEMS data and VPD, only the limited number of facilities that have been caught by manual inspection are labeled as positive samples, and the rest are unlabeled. This causes the aforementioned sparsity and missing label issues in the dataset, and makes the classification problem difficult.

Fortunately, PU learning (Bekker and Davis 2020), or learning from positive and unlabeled data, can overcome some of the challenges with our datasets. PU learning is an emerging machine learning method that has drawn considerable attention recently (Jaskie and Spanias 2019). In PU learning, each unlabeled sample could belong to either the positive or negative class. This fundamental

assumption is consistent with the context of environmental violator detection. The factories that are caught represent the labeled samples while others are the unlabeled samples.

We transform our dataset into the PU dataset as a triplet (X, Y, L) with $X = (X_1, X_2, \ldots, X_6)^\top \in \mathbb{R}^6$, a vector involving 6 data quality indicators, where Y is the true class label to present whether the factory is a violator, and L is a binary variable recording whether the factory was caught by MEEC. Obviously, the class label Y cannot be observed, but information about it can be derived from the value of L. For example, if the factory was caught by MEEC, then this factory must be a violator, which means $L = 1$ and implies

$$\mathbb{P}(Y = 1|L = 1) = 1.$$

However, if $L = -1$, we cannot know if the factory is a violator or not. That is, information on Y is missing.

PU learning methods are commonly divided into two categories according to labeling mechanism-based and labeling mechanism-free schemes (Bekker and Davis 2020). First, labeling mechanism-based PU learning generally treats the unlabeled samples as negative samples. We then build a decision function (classifier). Obviously this classifier is biased, as it assumes all unlabeled samples to be negative. To correct this bias, a labeling mechanism is built, which is incorporated into the training process. Second, labeling mechanism-free PU learning is generally built on the separability assumptions of the positive and negative samples. In other words, negative samples are very far from the positive samples in an unknown feature space. According to this separability property, reliable negative samples could be selected from the unlabeled samples. Thus, using supervised learning approaches on the labeled positive samples and selected reliable negative samples can obtain a decision function. In the next two subsections, we propose two kinds of PU learning approaches to the environmental violator detection problem, and show their effectiveness.

6.2 Labeling Mechanism-Based PU Learning for Violator Detection

The basic idea of labeling mechanism-based PU learning is to treat all unlabeled samples as negatives, and then design a new supervised learning strategy that can correct the learning bias due to the incorrect labeling, based on assuming the labeling mechanism is known. It should be mentioned that the labeled samples, which are in VPD, represent the factories that have been actually caught in a random inspection by MEEC. This randomness in inspection ensures that the labeling of facilities satisfies the *selected completely at random* (SCAR) assumption (Elkan and Noto 2008). That is,

$$\mathbb{P}(L = 1|X, Y = 1) = c,$$

which means any violators are caught completely at random, independent from their features. The value c is called label frequency, which denotes the probability of being caught. Based on this SCAR assumption, we can compute that

$$\mathbb{P}(L=1|X,Y=1)=\frac{\mathbb{P}(L=1,Y=1|X)}{\mathbb{P}(Y=1|X)}=\frac{\mathbb{P}(L=1|X)}{\mathbb{P}(Y=1|X)}=c.$$

Therefore, $\mathbb{P}(L=1|X)=c\mathbb{P}(Y=1|X)$. This formulation is critically important. It means that we can obtain $\mathbb{P}(Y=1|X)$ by the following steps. First, we estimate $\mathbb{P}(L=1|X)$, and then $\mathbb{P}(Y=1|X)$ can be corrected by using $\mathbb{P}(L=1|X)$ divided by constant c. Estimating $\mathbb{P}(L=1|X)$ is the common target of any supervised learning method that is trained with positive and unlabeled samples. The correction step is consistent with the insight of labeling mechanism-based PU learning. Based on this assumption, Zhang and Lee (2005) provide a new version of naive Bayes to handle the positive and unlabeled text data. Lee and Liu (2003) propose a weighted logistic regression. Elkan and Noto (2008) use a similar approach to obtain biased support vector machines. These methods are effective for their specific datasets. However, there has not been any justification of such methods. In this study, we propose a new PU learning framework and provide mathematical proof of its legitimacy for datasets that satisfy the SCAR assumption. Then, two new PU learning schemes are proposed.

Let us refer back to solving the binary classification problem with input and output pair (X,Y) by supervised learning approaches, where $X\in\mathcal{X}$ is a feature representation of a sample and $Y\in\{-1,1\}$ is the corresponding label. The aim of supervised learning for binary classification problems is to find a proper decision function (classifier) by minimizing the so-called *expected risk* (Bousquet et al. 2004), i.e.,

$$\min_f \mathbb{E}_{Y|X}[\ell(Y,f(X))|X], \tag{7}$$

where $\ell:\mathcal{X}\times\{-1,1\}\to\mathbb{R}$ is a well-defined loss function to quantify the misclassification error. Given X, it is natural to perform classification by comparing $\mathbb{P}(Y=1|X)$ and $\mathbb{P}(Y=-1|X)$, i.e., a reasonable decision function is

$$f(X)=sign(\mathbb{P}(Y=1|X)-\mathbb{P}(Y=-1|X)),$$

which is called a Bayesian classifier (or Bayesian decision boundary) of the binary classification problem (Lin 2004). According to Lin (2004) and Bartlett et al. (2006), a commonly accepted standard of the loss function ℓ in supervised learning is to require that the population minimizer of expected risk with the specific loss function results in the Bayesian classifier. That is to say that the minimizer of $\mathbb{E}_{Y|X}[\ell(f(X),Y)|X]$ has the same sign as $sign(\mathbb{P}(Y=1|X)-\mathbb{P}(Y=-1|X))$. The loss function that satisfies this unique property is called *Fisher consistent* (Lin 2004).

In this paper, we have shown that the violator detection problem is consistent with the PU learning framework due to the sparsity and missing label issues. However, a commonly accepted

way to handle the violator detection problem is to treat it as a classical binary classification problem, i.e., the unlabeled samples are considered as negative samples. In the following, we provide Theorem 1 to show that this naive approach can result in biased decision-making. Note that all the proofs in this paper are listed in Appendix A.

THEOREM 1. *Suppose that the SCAR labeling mechanism is satisfied for positive sample* $Y=1$ *and labeled sample* $L=1$*, i.e.,* $\mathbb{P}(L=1|X)=c\mathbb{P}(Y=1|X)$*, then the Bayesian classifier for the binary classification problem can be derived as*

$$f(X)=sign(\mathbb{P}(Y=1|X)-\mathbb{P}(Y=-1|X))=sign\Big(\frac{2-c}{c}\mathbb{P}(L=1|X)-\mathbb{P}(L=-1|X)\Big).$$

Theorem 1 actually implies that applying any Fisher consistent loss functions for the PU dataset directly leads to a large bias. We know that the optimal Bayesian classifier is $\mathbb{P}(Y=1|X)-\mathbb{P}(Y=-1|X)$ and Fisher consistent loss has the same sign as the Bayesian classifier. However, learning with a PU dataset directly results in the Bayesian classifier $\mathbb{P}(L=1|X)-\mathbb{P}(L=-1|X)$, which is not equal to $\mathbb{P}(Y=1|X)-\mathbb{P}(Y=-1|X)$ according to the result of Theorem 1. Therefore, Theorem 1 provides the theoretical reason why directly training a classifier on a PU dataset is biased. Note that Elkan and Noto (2008), Lee and Liu (2003), and Zhang and Lee (2005) all use the sample weighting strategy to correct the learning bias. However, they were not able to explain the theoretical reason for such corrections. Sharing the same insight, we would like to propose a new weighting strategy to correct the learning bias in PU learning. The fundamental idea of this strategy is to use the result of Theorem 1 to construct a weighting function that can transfer some loss functions to be Fisher consistent.

Let us denote a weighting function

$$g(Y)=\begin{cases}2-c, & \text{if } Y=1,\\ c, & \text{if } Y=-1.\end{cases} \tag{8}$$

It is well known that the loss function (logistic loss) for logistic regression is $\ell(Y,f(X))=\log(1+\exp(-Yf(X)))$, and the loss function (exponential loss) of AdaBoost is $\ell(Y,f(X))=\exp(-Yf(X))$. Both are Fisher consistent (Friedman et al. 2000). The next two theorems show how to correct the above two losses to be Fisher consistent for PU learning by using weighting function $g(Y)$. For the logistic regression,

THEOREM 2. *Suppose that the SCAR labeling mechanism is satisfied, i.e.,* $\mathbb{P}(L=1|X)=c\mathbb{P}(Y=1|X)$*, and further denote* $f_1^*(X)=\arg\min\mathbb{E}_{L|X}[g(L)\ell(f(X),L)|X]$ *where* $\ell(f(X),L)=\log(1+\exp(-Lf(X)))$*, then we have*

$$sign(f_1^*(X))=sign(\mathbb{P}(Y=1|X)-\mathbb{P}(Y=-1|X)).$$

For the AdaBoost,

THEOREM 3. *Suppose that the SCAR labeling mechanism is satisfied, i.e.,* $\mathbb{P}(L=1|X) = c\mathbb{P}(Y=1|X)$, *and further denote* $f_2^*(X) = \arg\min \mathbb{E}_{L|X}[g(L)\ell(f(X),L)|X]$ *where* $\ell(f(X),L) = \exp(-Lf(X))$, *then*

$$sign(f_2^*(X)) = sign(\mathbb{P}(Y=1|X) - \mathbb{P}(Y=-1|X)).$$

Theorems 2 and 3 show that incorporating the weighting function $g(Y)$ into the training stage of logistic regression and AdaBoost can correct the decision bias that comes from the PU dataset if the SCAR assumption is satisfied. In practice, the real distribution of the PU dataset cannot be obtained, so we have to use the empirical risk minimization (9) (Bousquet et al. 2004) to take the place of expected risk minimization (7) in the training stage. Then we propose the following algorithms for the PU logistic regression and AdaBoost. To implement Algorithm 1, we need to set the parameter c in the weighting function $g(Y)$. The tuning of the parameter will be addressed in Section 7.

Algorithm 1 Labeling Mechanism-Based PU Learning Algorithm for Violator Detection

1: Transform $\mathbf{R}$ into $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times 6}$ by the proposed data quality assessment framework, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{i6})^\top, i = 1, \ldots, n$ is the data quality indicators of the i-th factory.

2: Compute

$$\hat{f} \in \min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} g(l_{it_i}) \ell(f(\mathbf{x}_i), l_{it_i}) \right\}, \tag{9}$$

where $g(L)$ is defined by (8), and $\ell(f(X),L) = \exp(-Lf(X))$ and $\ell(f(X),L) = \log(1+\exp(-Lf(X)))$ are used for logistic regression and AdaBoost, respectively.

6.3 Labeling Mechanism-Free PU Learning for Violator Detection

The separability assumption is the fundamental basis of labeling mechanism-free PU learning. This assumption considers that all the positive samples are very similar to the labeled samples and the negative samples are very different from them in an unknown feature space (Bekker and Davis 2020). Based on this assumption, PU learning can be formulated as a two-step technique as follows.

- Step 1: Constructing a reasonable approach to identify reliable negative samples from the unlabeled samples in a feature space.
- Step 2: Applying supervised learning techniques with the positive and reliable negative samples to obtain the decision function (classifier).

In the first step, the key objective is to define a proper feature space so that positive and negative samples in the new feature space are clearly separated, and then the unlabeled samples, which are far away from labeled samples, are selected as reliable negatives. Thus, how to define a proper feature space that leads to the separability assumption becomes a crucial problem in labeling

mechanism-free PU learning. In the existing PU learning literature (Chaudhari and Shevade 2012, Fung et al. 2005, Liu and Peng 2014), this problem is handled by applying domain knowledge. For example, many PU learning approaches have addressed text classification problems, therefore many feature engineering techniques, such as TFIDF (Term Frequency Inverse Positive-Negative Document Frequency) and word embedding, provide a new feature space for the texts. In this situation, a natural way to select negative samples is to use a clustering technique (Chaudhari and Shevade 2012, Fung et al. 2005, Li and Liu 2003, Liu and Peng 2014, Lu and Bai 2010), such as k-means clustering, because in the feature space, the positive and negative samples are assumed to be separable.

In the second step, the labeled positive and reliable negative samples are employed to train a decision function. Obviously, any supervised learning method, such as support vector machines (Li and Liu 2003), logistic regression (Lee and Liu 2003), and naive Bayes (Lu and Bai 2010), could be used in this context. Finally, to obtain more reliable negative samples and to enhance the robustness of the classifier, previous research also apply an iterative strategy (Li and Liu 2003). In this iterative strategy, researchers utilize the obtained classifiers to classify the remaining unlabeled data, in the hope of providing more reliable negative samples. This results in an updated classifier. This iterative process repeats itself until a satisfactory result is obtained.

Using a similar insight, we propose the following labeling mechanism-free PU learning algorithm (Algorithm 2) for violator detection. Given raw CEMS data $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$ and VPD $\mathbf{l} = (l_{1t_1}, \ldots, l_{nt_n})$ where $l_{it_i} = 1$ means the i-th factory is punished at timestamp t_i, we denote factory index set as $\mathcal{I} = \{1, 2, \ldots, n\}$, labeled positive sample index set as $\mathcal{L} = \{i : l_{it_i} = 1\}$, and unlabeled sample index set as $\mathcal{U} = \{i : l_{it_i} = -1\}$, thus $\mathcal{I} = \mathcal{L} \cup \mathcal{U}$. Given any index set $\mathcal{M}$ and its cardinality of m, we further denote $\mathbf{X}_{\mathcal{M}} = (\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_m})$ as the sub-matrix of $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\top$, $\mathbf{x}_i$ as the i-th row vector, and the index as $i_j \in \mathcal{M}, j = 1, \ldots, m$. Algorithm 2 is based on the above formulation.

Algorithm 2 provides a new labeling mechanism-free PU learning method for violator detection. Comparing with existing approaches, this new method incorporates the feature space supported by the data quality assessment framework, which is constructed by the domain knowledge of SMEs. Afterwards, we utilize k-means clustering and an iterative scheme to obtain reliable negative samples. The effectiveness of these techniques has been validated in a number of studies (Chaudhari and Shevade 2012, Fung et al. 2005, Li and Liu 2003, Liu and Peng 2014, Lu and Bai 2010).

7 PU Learning Models and the Detection of Environmental Violators

7.1 Data Processing

In this section, we describe an application of the proposed techniques to the integrated CEMS dataset and VPD. The VPD contains 372 records of factories caught through manual inspection

Algorithm 2 Labeling Mechanism-Free PU Learning Algorithm for Violator Detection

Given the maximal clustering number K and iterative number J. Transform $\mathbf{R}$ into $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times 6}$ by the proposed data quality assessment framework, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{i6})^\top, i = 1, \ldots, n$ is the data quality indicators of the i-th factory.
for $k = 2 : K$ **do**
    Apply k-means clustering algorithm on data quality matrix $\mathbf{X}$.
    Compute the average distances d_k of the samples that are in the cluster which does not involve positive samples, with all positive samples.
end for
Select $k^* = \arg\max_{k=1:K} d_k$.
Run K-means via the clustering number k^* and select the samples that are in the cluster which does not involve the positive samples as reliable negatives, and denote the selected reliable negative samples with the index set $RN_1 = \{i : \text{the } i\text{-th sample is selected as reliable negative sample}\}$.
for $j = 1 : J$ **do**
    Use a supervised learning approach on the dataset $\{(\mathbf{x}_j, l_{jt_j}) : j \in \mathcal{L} \cup RN_j\}$ to obtain a decision function f_j. Apply f_j on the sample set with index $\mathcal{I} - (\mathcal{L} \cup RN_j)$ to predict their labels.
    Add the predicted negative samples into RN_j to obtain a new reliable negative sample set.
end for
Output: f_J as the final classifier.

during 2016-2017. Of these records, 118 were caught for gas violation. The set consisting of the gas violation factories that have been caught is labeled as the sample index set $\mathcal{L}$. Out of the 7,643 facilities, there are 3,482 relevant facilities that installed a gas emission monitoring system. We remove the 118 facilities from these 3,482 facilities and select the remaining ones into the unlabeled set $\mathcal{U}$. Thus, $\mathcal{L}$ and $\mathcal{U}$ are disjoint.

Before building predictive models, we process the data from CEMS and VPD to ensure appropriate integration of the two datasets. First, we examine the date that a particular facility has been caught, and remove entries in the CEMS dataset post that date (note that although VPD is published quarterly, it does specify the date when a particular facility has been caught). Since it is reasonable to expect the violating facility to change its reporting behavior after it has been caught, we must remove the post-caught data to ensure the consistency of the data used to predict environmental predictors. Second, we use the remaining dataset to compute data quality indicators. Note that these indicators are ratio data. For example, the first indicator, abnormal ratio, is computed by dividing accumulated abnormal period by accumulated running time. Therefore, the computation of the indicators is time scale free. Thus, the removal of post-caught records, as reported in step one, will not impact the computation of the ratio data. Third, certain facilities may contain several outlets. When we calculate the reporting data quality indicators, the average value of outlets is utilized as the reporting data quality indicators for the facility.

7.2 Model Description

Table 3 provides a summary of the models we built. We now provide a detailed discussion of these. We built five types of model: (i) learning from raw CEMS data of $\mathcal{L} \cup \mathcal{U}$, using Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), which is a specific recurrent neural

network (RNN) architecture that was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs; (ii) learning from $\mathcal{L}\cup\mathcal{U}$ with the 6 data quality features, using logistic regression and AdaBoost; (iii) learning from $\mathcal{L}\cup\mathcal{U}$ with the 6 data quality features using the algorithm proposed by Lee and Liu (2003) (denoted as WLR-PU); (iv) learning from $\mathcal{L}\cup\mathcal{U}$ with the 6 data quality features using Algorithm 1; and (v) learning from $\mathcal{L}\cup\mathcal{U}$ with the 6 data quality features using Algorithm 2. We also test a naive rule-based model based on insights from the qualitative interviews. We report the results of the rule-based model in Appendix B. The results are inferior to other prediction models presented here. Tables 4 and 5 provide summaries of model results.

Table 3 Model Comparison

Method	Raw Data or Features	PU Learning	Evaluation Metric
LSTM	Raw Data	No	AUC, precision, recall, F1-score
Logit	Features	No	AUC, precision, recall, F1-score
AdaBoost	Features	No	AUC, precision, recall, F1-score
WLR-PU	Features	Yes	recall, F_{pu}
PU Logistic	Features	Yes	recall, F_{pu}
PU AdaBoost	Features	Yes	recall, F_{pu}
PU-2 Logistic	Features	Yes	AUC, precision, recall, F1-score
PU-2 AdaBoost	Features	Yes	AUC, precision, recall, F1-score

Scenario (i) is the baseline: learning a classifier from raw CEMS data, characterized by irregularity and outliers. To deal with the irregularity, we adopt the state-of-the-art deep learning technique LSTM. Scenario (ii) employs classical supervised learning approaches for the CEMS data that we recognized as a fully labeled dataset. This represents a scenario that utilizes feature engineering but does not utilize PU learning, i.e., we treat l_{it_i} values as the true labels of CEMS data and the unlabeled samples as negative samples. Note that l_{it_i} values lead to an unbalanced training set because factories that are caught are relatively smaller in number than factories that are not caught. Then we use the classical sample weighting method (He and Garcia 2009) to train the logistic regression and AdaBoost with the loss formulation as:

$$\min_f \left\{ \frac{n^+ + n^-}{n^+} \sum_{l_{it_i}=1} \ell(f(\mathbf{x}_i), l_{it_i}) + \frac{n^+ + n^-}{n^-} \sum_{l_{it_i}=-1} \ell(f(\mathbf{x}_i), l_{it_i}) \right\}, \tag{10}$$

where $\ell(f(X), L) = \exp(-Lf(X))$ and $\ell(f(X), L) = \log(1 + \exp(-Lf(X)))$ are used for logistic regression and AdaBoost respectively, and n^+, n^- are the number of labeled samples and unlabeled samples. To make comparison with other PU learning algorithms, we also include scenario (iii). Even though many PU learning classification techniques have been proposed over the last decade, most of them focus on the classification of text data, which makes them inapplicable to our research context. We choose WLR-PU, which is the most suitable technique for our CEMS application. This

Table 4 Mean and Standard Deviation of AUC, Precision, Recall, F_1-score, and F_{pu} for Different Comparable Models

Method	AUC	Precision	Recall	F_1	F_{PU}
LSTM	0.57(0.08)	0.04(0.01)	0.93(0.03)	0.07(0.01)	-
Logit	0.71(0.07)	0.03(0.006)	0.73(0.14)	0.07(0.012)	-
AdaBoost	0.80(0.08)	0.10(0.03)	0.46(0.15)	0.16(0.05)	-
WLR-PU	0.78(0.06)	0.03(0.006)	0.69(0.15)	0.06(0.01)	1.66(0.64)
PU Logistic	0.75(0.06)	0.03(0.006)	0.73(0.17)	0.06(0.01)	1.47(0.47)
PU AdaBoost	0.80(0.08)	0.04(0.01)	0.76(0.14)	0.07(0.01)	1.83(0.47)
PU-2 Logistic($J=1$)	0.97(0.03)	0.53(0.18)	0.87(0.10)	0.65(0.14)	14.99(4.80)
PU-2 AdaBoost($J=1$)	0.96(0.05)	0.94(0.10)	0.83(0.14)	0.87(0.10)	29.44(5.86)
PU-2 Logistic($J=5$)	0.96(0.04)	0.44(0.16)	0.86(0.11)	0.57(0.13)	14.28(5.08)
PU-2 AdaBoost($J=5$)	0.94(0.06)	0.88(0.15)	0.79(0.15)	0.82(0.13)	30.42(7.21)
PU-2 Logistic($J=9$)	0.96(0.03)	0.44(0.16)	0.86(0.11)	0.56(0.13)	14.13(4.92)
PU-2 AdaBoost($J=9$)	0.94(0.06)	0.83(0.21)	0.79(0.15)	0.80(0.17)	29.51(9.03)

Table 5 Mean and Standard Deviation of Recall@k for Different Comparable Models

k	LSTM	B-Logit	B-Ada	WLR-PU	PU Log	PU Ada	PU2Log(1)	PU2Ada(1)	PU2Log(5)	PU2Ada(5)	PU2Log(9)	PU2Ada(9)
10	0.03(0.01)	0.09(0.05)	0.34(0.13)	0.14(0.1)	0.19(0.11)	0.34(0.13)	0.21(0.01)	0.2(0.02)	0.21(0.0)	0.19(0.04)	0.21(0.0)	0.19(0.04)
20	0.03(0.01)	0.1(0.07)	0.37(0.13)	0.2(0.11)	0.27(0.12)	0.37(0.13)	0.41(0.03)	0.39(0.04)	0.41(0.02)	0.3(0.07)	0.41(0.01)	0.29(0.06)
30	0.06(0.01)	0.1(0.06)	0.4(0.13)	0.25(0.13)	0.31(0.14)	0.41(0.14)	0.58(0.06)	0.56(0.07)	0.58(0.06)	0.36(0.1)	0.56(0.06)	0.35(0.09)
40	0.09(0.02)	0.11(0.06)	0.43(0.14)	0.28(0.13)	0.31(0.14)	0.44(0.15)	0.69(0.11)	0.7(0.12)	0.69(0.1)	0.42(0.14)	0.65(0.1)	0.42(0.15)
50	0.10(0.03)	0.11(0.07)	0.45(0.14)	0.3(0.14)	0.32(0.14)	0.46(0.15)	0.73(0.1)	0.78(0.14)	0.73(0.09)	0.47(0.18)	0.7(0.09)	0.48(0.2)
60	0.12(0.05)	0.11(0.07)	0.48(0.14)	0.32(0.14)	0.35(0.14)	0.47(0.16)	0.77(0.08)	0.81(0.14)	0.77(0.07)	0.52(0.21)	0.75(0.07)	0.52(0.22

scenario can be viewed as a baseline PU learning algorithm. Finally, scenarios (iv) and (v) represent our proposed methods, where scenario (iv) represents labeling mechanism-based PU learning and (v) represents labeling mechanism-free PU learning. We should mention that the unknown labeling frequency c is treated as a tuning parameter, which is selected by 5-folder cross validation in Algorithm 1. In Algorithm 2, we set the maximal cluster number $K = 30$ and iterative number J from 1 to 9.

In each scenario, we split the whole dataset into a training set with 80% of samples and a testing set with 20% of samples and repeat the experiment 100 times. The average and standard deviation of AUC, precision, recall, F1-score, F_{pu}, and recall@k, which are the commonly used measurements in supervised learning and PU learning, are reported in Tables 4 and 5. Because different scenarios represent different model assumptions, the model fit indicators should be interpreted in the context of the type of model being evaluated. We provide more detail on this in the next subsection.

7.3 Model Evaluation and Results

First of all, scenarios (i) (i.e., LSTM) and (ii) (i.e., logistic regression and AdaBoost) represent models without the PU learning framework. In these models, the labeled samples are treated as positive samples and nonlabeled samples are treated as negative samples. In these three models, the violator detection problem is treated as a classical binary classification problem, and as such AUC, precision, recall, and F1-score can be utilized to compare their performance. From Table 4,

the AUC value of LSTM is less than that of logistic regression or AdaBoost. This demonstrates the advantage of the proposed data quality features over raw data, thereby benefiting the detection of falsification issues in reported data. However, the precision and F1-scores across all three models are very poor due to the unlabeled dataset, i.e., these models treat the unlabeled data as true negatives, which is incorrect. This justifies the need to employ the proposed PU learning framework in the subsequent scenarios.

Second, we compare the results of violator detection under the PU learning framework by WLR (Lee and Liu 2003) and our proposed PU logistic regression and AdaBoost. Note that although we report AUC, precision, and F1-scores in Table 4, they are not the best indicators of model fit due to the existence of unlabeled data. Along the same line, we do not report precision@k and f@k because they require clearly labeled datasets. Under the PU learning framework, recall and F_{pu} are two meaningful measurements (Lee and Liu 2003). Recall measures the accuracy of catching the true offenders, and can be denoted as $r = \mathbb{P}(\hat{Y} = 1|Y = 1)$. Under the SCAR assumption, the recall can be estimated from PU data as $r = \mathbb{P}(\hat{Y} = 1|L = 1)$. Based on the definition of recall, we know that higher recall value means higher capability to catch violators. $\mathrm{F}_{pu} = \frac{r^2}{\mathbb{P}(\hat{Y}=1)}$, which has been proven to have the same property as the F1-score, could be estimated from PU data, and the F_{pu} indicates efficiency in catching violators (Bekker and Davis 2020). From Table 4, the PU AdaBoost shows high accuracy in catching violators (best recall), while at the same time, it uses less resources (best F_{pu} compared to other PU learning models).

Third, we present the AUC, precision, recall, and F1-scores for the two labeling mechanism-free PU learning methods, which are denoted as PU2Ada and PU2Log, for the use of AdaBoost and logistic regression in Step 9 of Algorithm 2. We tried different iterative numbers J from 1 to 9 and found that the optimal number of cluster k^* is always around 18. For different Js, the results show that the algorithms are robust and $J = 1$ can achieve the best recall value. This means that the data quality indicators we proposed fit the separability assumption on the CEMS data and VPD, because using k-means algorithms once suffices to separate the labeled (positive) and reliable negative samples.

Finally, we report the recall@k for different $k = 10, 20, \ldots, 60$, because recall is the only metric that can be used to compare the capability of catching violators for all comparable methods. Based on the results of Table 5, the labeling mechanism-free PU learning (PU2Ada and PU2Log) still has the best performance in catching violators. In summary, the labeling mechanism-free PU learning algorithms show signs of robustness and good performance in detecting violators, due to their simplicity, efficiency, and accuracy in detection.

7.4 Robustness Test

Our key models are built on waste gas datasets. In order to test the robustness and to enhance the generalizability of our model, we have also applied our prediction models to wastewater datasets. Wastewater data collection uses an entirely different mechanism, and collects data on an entirely different set of substances at completely different intervals. However, there is some preliminary evidence that our predictive models may work well for the wastewater context. The results of predictive models based on this context are provided in Appendix C. We note that due to the drastically different context, some parameters may need to be fine-tuned accordingly. We call for future research to investigate this further.

7.5 Field Test

The previous sections have demonstrated the predictive power of the proposed CEMS data quality assessment framework and the PU learning methods based on the empirical data we collected. Next, we take advantage of the findings from our prediction models and prescribe a new means of violation detection, which we then use in a field test. This field test provides evidence of the usefulness of our newly prescribed process. It also serves as a means to further validate the results of our prediction model.

We prescribe four steps to systematically generate a list of suspected environmental violators. These four steps are documented in Figure 2. First, we extract raw CEMS data from July 1, 2017 to December 31, 2017 in one of the provinces in China. Second, we use the data quality assessment framework to compute a set of data quality indicators for each of the reporting facilities. Third, we run the labeling mechanism-free PU logistic regression model that is obtained in Algorithm 2 with this set of indicators, and compute the violation status. Finally, we rank firms based on predicted violation probability and obtain a list of the top five suspicious facilities. We provided the list to MEEC, which subsequently assembled a team to inspect these five suspicious factories in 2018. We report the results based on this targeted inspection mechanism below.

MEEC inspected five case companies. Case Company A is a large steel-making company. Its main source of pollution is air pollution. In this company, gas pollutants are discharged through the chimney after being treated. The inspection staff found a hidden break at the bottom of the chimney used to discharge gas pollutants (see Figure 3). This break served to introduce more oxygen to dilute the gas pollutants, thereby distorting the readings on the monitoring devices.

Case Companies B and C changed their monitoring devices immediately before the arrival of staff to perform the facility inspection, indicating that perhaps there was a leak of information to these facilities, which prompted them to switch the devices. According to our SMEs, it is highly

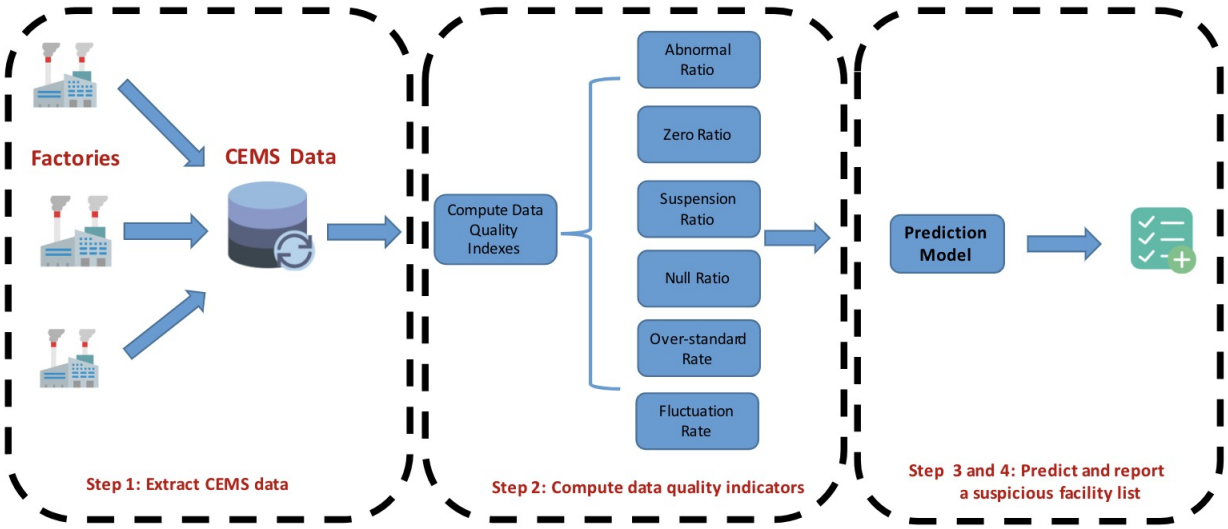

Figure 2 Prescribed New Process of Targeted Inspection

Figure 3 Evidence of Tampering with Monitoring Device

likely that these facilities would have been caught circumventing the monitoring devices, had the devices not been changed right before the inspection. The staff were not able to find any evidence of tampering with the device in Case Company D. However, the company cannot explain the data quality problems.

Finally, Company E showed evidence of introducing new pollution control equipment for environmental protection. The new equipment had recently reduced the average concentration of pollution (between the time of our analysis and the inspection). Note that the difference between Case Companies B and C, and Case Company E is that the latter showed evidence of improving production and waste control process without changing the monitoring device, while the former two case companies showed no evidence of improving the production or waste control process; at the same time, they showed evidence of changing the monitoring device immediately before the inspections. Along the same lines, the difference between Case Company D and Case Company E is that the former showed no evidence of improving the production or waste control process and could not explain the

abnormal readings, while the latter showed improvements in the production and waste control process, which provided a logical explanation for the improvement of readings at the time of inspection.

8 Conclusions and Managerial Implications

Environmental violations pose severe threats to the triple bottom line (e.g., Lee and Tang 2018, Mendelssohn et al. 2012) and distort fair competition for non-violating firms (Delmas and Keller 2005). However, low fidelity in reported environmental monitoring data is not uncommon and impedes public agencies ability to detect hidden violators. We conduct a study to make predictions on environmental violators. Along the way, we showcase a holistic approach to solve a practical problem in sustainability enforcement with descriptive, predictive, and prescriptive data analytics techniques. Our approach has been tested in a field test and the results are satisfactory and encouraging.

8.1 Contributions

Our study contributes to both sustainability literature and big data analytics. First, we build a reporting data quality assessment framework, which provides a useful way to measure the level of fidelity in the reported environmental monitoring data. Its significance lies beyond the detection of inaccuracy in data caused by the noise in big data (Dey and Kumar 2013), in that the framework can detect low fidelity caused by intentional manipulation of the data. As far as we know, we are the first to derive such a measurement scale, which can serve as an excellent tool for public offices to deploy in situations where there are conflicts between government reporting requirements and the internal efficiency of firms (Marquis and Qian 2014).

Second, we utilize approaches to build prediction models of environmental violators and prescribe a new process to catch cheaters. These prediction models and the new process greatly increase the efficiency and effectiveness of environmental agencies' operations. Third, while building prediction models, we develop two new PU learning techniques to overcome the sparsity and missing label challenges in the empirical dataset. The usability of these two new PU learning techniques we developed is not restricted to our unique dataset, but can be generalized to other datasets where sparsity and mislabeled data exist in a binary classification problem. Fourth, our research answers calls from operations and supply chain management scholars to use a holistic approach (Swaminathan 2018, Wang et al. 2016) to solve operational problems. Our study is one of the first to integrate descriptive, predictive, and prescriptive data analytics elements to solve a pressing problem in sustainable operations.

Our research has great practical implications. Public agencies can utilize the prediction model to compute the probability of environmental violation for each reporting facility. This is the most

direct and effective mechanism to identify violators, and potentially reduces the resource investment associated with manual inspections. In addition, once the improved detection algorithm has been fully implemented and communicated to pollutant-producing facilities, we expect more facilities will report high quality data and proactively work on improving their internal operations to reduce the severity of pollutants.

Although we studied the detection of hidden environmental violators using datasets from China, we expect that the research findings can be easily adaptable to environmental protection in other countries. For example, the EPA in the United States utilizes the same CEMS devices to monitor environmental pollutants, and can benefit from pioneering the use of our data quality assessment framework to measure the quality of reported waste gas data. The resulting data quality indices can then be used to predict the under-reporting of emissions by manufacturing facilities in the US. Our predictive model shows superb predictive power and thus can be used in conjunction with the control chart approach that the EPA is rolling out (EPA 2016). Similarly, we could apply our model to data collected on oilsands operations in Canada and evaluate whether data falsification is involved in the reporting of pollutant data.

8.2 Future Research Directions

This research can be extended in several directions. First, the VPD is limited in the number of records compared to the size of the CEMS dataset. However, the sample size is increasing with the passing of time. With a larger sample size, future research can utilize other complex predictive techniques in machine learning to refine the predictive model and enhance prediction accuracy. Second, our research focuses on the detection of violations for gas pollutants. We also validate its usefulness in the wastewater context. Future research can replicate our research for the detection of other types of pollutants. Such research will enhance the generalizability of the reporting data quality assessment framework and potentially shed light on the detection of anomalies in data reported by other Internet of Things devices. Lastly, our research is data-driven. We made it clear that our objective is prediction, not finding causal relationships (Kitchin 2014). Future research can take advantage of our data-driven findings and build theoretical models. This type of follow-up research will help us explain and understand why facilities follow and/or violate environmental regulations.

Appendix

A Proofs

A.0.1 Proof of Theorem 1 *According the definition of SCAR, we know that* $\mathbb{P}(L=1|X)=c\mathbb{P}(Y=1|X)$ *and* $\mathbb{P}(L=-1|Y=-1,X)=1$*, so we can compute that*

$$\mathbb{P}(L=-1|X)=\mathbb{P}(L=-1|Y=-1,X)\mathbb{P}(Y=-1|X)+\mathbb{P}(L=-1|Y=1,X)\mathbb{P}(Y=1|X)$$

$$= \mathbb{P}(Y=-1|X) + (1-c) * \mathbb{P}(Y=1|X)$$
$$= \mathbb{P}(Y=-1|X) + \frac{1-c}{c} * \mathbb{P}(L=1|X),$$

and

$$\mathbb{P}(Y=-1|X) = \mathbb{P}(L=-1|X) - \frac{1-c}{c} * \mathbb{P}(L=1|X).$$

Thus,

$$\begin{aligned}\mathbb{P}(Y=1|X) - \mathbb{P}(Y=-1|X) &= \frac{1}{c}\mathbb{P}(L=1|X) - \mathbb{P}(L=-1|X) + \frac{1-c}{c}\mathbb{P}(L=1|X) \\ &= \frac{2-c}{c}\mathbb{P}(L=1|X) - \mathbb{P}(L=-1|X).\end{aligned}$$

This implies the results.

A.0.2 Proof of Theorem 2 *Let us compute and denote that*

$$\begin{aligned}G(f(X)) &= \mathbb{E}_{L|X}(g(L)\exp\{-Lf(X)\}) \\ &= (2-c)\mathbb{P}(L=1|X)\log(1+\exp\{-f(X)\}) + c\mathbb{P}(L=-1|X)\log(1+\exp\{f(X)\}).\end{aligned}$$

.

Therefore, f_1^* must satisfy

$$\frac{\partial G(f(X))}{\partial f(X)} = -\frac{(2-c)\mathbb{P}(L=1|X)}{1+\exp\{f(X)\}} + \frac{c\mathbb{P}(L=-1|X)\exp\{f(X)\}}{1+\exp\{f(X)\}} = 0,$$

and we can obtain

$$f_1^*(X) = \log\frac{(2-c)\mathbb{P}(L=1|X)}{c\mathbb{P}(L=-1|X)}.$$

Thus,

$$\begin{aligned}sign(f_1^*(X)) &= sign\Big(\log\frac{(2-c)\mathbb{P}(L=1|X)}{c\mathbb{P}(L=-1|X)}\Big) \\ &= sign\Big(\frac{2-c}{c}\mathbb{P}(L=1|X) - \mathbb{P}(L=-1|X)\Big) \\ &= sign(\mathbb{P}(Y=1|X) - \mathbb{P}(Y=-1|X)),\end{aligned}$$

where the last equation is based on Theorem 1.

A.0.3 Proof of Theorem 3 *Let us compute and denote that*

$$G(f(X)) = \mathbb{E}_{L|X}(g(L)\exp\{-Lf(X)\}) = (2-c)\mathbb{P}(L=1|X)\exp\{-f(X)\} + c\mathbb{P}(L=-1|X)\exp\{f(X)\}.$$

Therefore, f_2^* must satisfy

$$\frac{\partial G(f(X))}{\partial f(X)} = -(2-c)\mathbb{P}(L=1|X)\exp\{-f(X)\} + c\mathbb{P}(L=-1|X)\exp\{f(X)\} = 0,$$

and we can obtain

$$f_2^*(X) = \frac{1}{2}\log\frac{(2-c)\mathbb{P}(L=1|X)}{c\mathbb{P}(L=-1|X)}.$$

Thus,

$$\begin{aligned} sign(f_2^*(X)) &= sign\Big(\log\frac{(2-c)\mathbb{P}(L=1|X)}{c\mathbb{P}(L=-1|X)}\Big) \\ &= sign\Big(\frac{2-c}{c}\mathbb{P}(L=1|X)-\mathbb{P}(L=-1|X)\Big) \\ &= sign(\mathbb{P}(Y=1|X)-\mathbb{P}(Y=-1|X)), \end{aligned}$$

where the last equation is based on Theorem 1.

B Naive Check

In Section 5, we put forward the CEMS data quality assessment framework based on domain knowledge from SMEs. We test a naive rule-based model based on the qualitative interviews we conducted with the SMEs.

We use the same data processing method that is described in Section 7 to construct the training set from CEMS data and VPD. We split the whole dataset into a training set (80% of samples) and a testing set (20% of samples) and repeat the experiment 100 times. We use the null ratio as a prediction rule, and its average ROC curve is shown in Figure 4. Figure 4 shows that the naive rule-based method cannot achieve the desired prediction power. The average AUC of the null ratio is 0.64. Compared with the logistic regression and/or AdaBoost in scenario (i) of Section 7, the prediction power is relatively low. This provides motivation to investigate a more powerful method to detect violators.

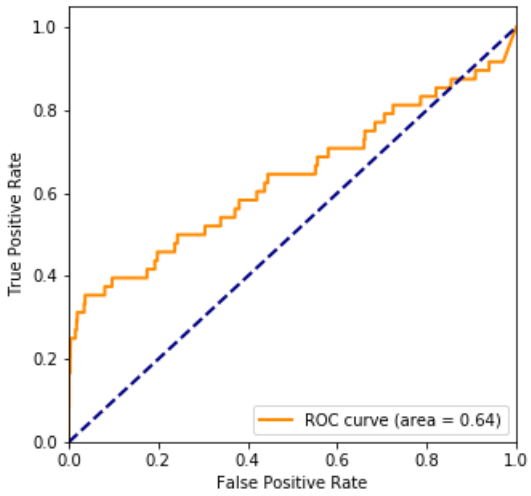

Figure 4 ROC Curve of Rule Derived by Null Ratio.

Table 6 Mean and Standard Deviation of AUC, Precision, Recall, F1-score, and F_{pu} for Different Comparable Models for Wastewater Dataset

Method	AUC	Precision	Recall	F_1	F_{PU}
LSTM	0.68(0.01)	0.04(0.0)	0.74(0.08)	0.07(0.0)	-
Logit	0.54(0.06)	0.02(0.006)	0.35(0.14)	0.034(0.01)	-
AdaBoost	0.78(0.07)	0.08(0.01)	0.56(0.12)	0.14(0.02)	-
WLR-PU	0.58(0.06)	0.018(0.005)	0.38(0.13)	0.035(0.01)	0.53(0.32)
PU Logistic	0.57(0.06)	0.013(0.0003)	0.98(0.04)	0.027(0.0006)	0.97(0.058)
PU AdaBoost	0.78(0.07)	0.037(0.01)	0.70(0.12)	0.07(0.017)	1.86(0.57)
PU-2 Logistic($J=1$)	1.0(0.0)	0.71(0.37)	0.89(0.08)	0.72(0.31)	8.14(4.48)
PU-2 AdaBoost($J=1$)	0.99(0.01)	0.71(0.21)	0.86(0.08)	0.76(0.15)	9.56(2.86)
PU-2 Logistic($J=5$)	0.99(0.0003)	0.70(0.36)	0.88(0.08)	0.71(0.30)	7.98(4.19)
PU-2 AdaBoost($J=5$)	0.90(0.02)	0.51(0.27)	0.80(0.08)	0.57(0.22)	9.62(4.94)
PU-2 Logistic($J=9$)	1.0(0.0)	0.70(0.36)	0.89(0.08)	0.71(0.30)	7.97(4.19)
PU-2 AdaBoost($J=9$)	0.85(0.03)	0.45(0.26)	0.78(0.08)	0.52(0.23)	9.01(5.07)

C Prediction Results for Wastewater Dataset

Table 7 Mean and Standard Deviation of Recall@k for Different Comparable Models for Wastewater Data

k	LSTM	B-Logit	B-Ada	WLR-PU	PU Log	PU Ada	PU2Log(1)	PU2Ada(1)	PU2Log(5)	PU2Ada(5)	PU2Log(9)	PU2Ada(9)
10	0.08(0.01)	0.01(0.02)	0.39(0.09)	0.01(0.02)	0.02(0.03)	0.4(0.08)	0.1(0.01)	0.11(0.0)	0.1(0.01)	0.11(0.0)	0.09(0.02)	0.11(0.01)
20	0.16(0.05)	0.01(0.02)	0.46(0.11)	0.01(0.02)	0.03(0.04)	0.46(0.09)	0.18(0.02)	0.21(0.01)	0.18(0.03)	0.21(0.01)	0.18(0.03)	0.2(0.02)
30	0.20(0.03)	0.01(0.02)	0.48(0.11)	0.02(0.03)	0.03(0.04)	0.49(0.09)	0.24(0.05)	0.32(0.03)	0.25(0.06)	0.32(0.02)	0.25(0.06)	0.3(0.04)
40	0.22(0.06)	0.01(0.02)	0.49(0.11)	0.02(0.03)	0.04(0.04)	0.5(0.1)	0.3(0.08)	0.42(0.03)	0.32(0.08)	0.42(0.03)	0.31(0.09)	0.4(0.06)
50	0.22(0.07)	0.01(0.02)	0.5(0.11)	0.03(0.04)	0.05(0.04)	0.51(0.1)	0.36(0.1)	0.52(0.04)	0.39(0.1)	0.52(0.05)	0.38(0.1)	0.49(0.06)
60	0.22(0.07)	0.01(0.02)	0.51(0.11)	0.04(0.04)	0.05(0.05)	0.52(0.1)	0.41(0.1)	0.63(0.04)	0.45(0.1)	0.61(0.06)	0.44(0.11)	0.57(0.06)

Acknowledgments

The authors gratefully acknowledge kindly and useful comments and suggestions from the department editor, associate editor, and two anonymous reviewers. The work of Xiangyu Chang was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 11771012 and 61502342. The work of Xin Bo was supported in part by the National Key Research and Development Program of China 2019YFE0194500.

All the URLs are last accessed on Sep 14, 2020.

References

AN (2019) Ford facing U.S. criminal probe into emissions testing. *Automotive News*. URL https://www.autonews.com/regulation-safety/ford-facing-us-criminal-probe-emissions-testing.

Arenas D, Rodrigo P (2016) On firms and the next generations: Difficulties and possibilities for business ethics inquiry. *Journal of Business Ethics* 133(1):165–178.

Ballou DP, Pazer HL (1985) Modeling data and process quality in multi-input, multi-output information systems. *Management Science* 31(2):150–162.

Bartlett P, Jordan M, McAuliffe J (2006) Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473):138–156.

Bekker J, Davis J (2020) Learning from positive and unlabeled data: A survey. *Machine Learning* 109(4):719–760.

Bernard Y, Godard L, Zouaoui M (2018) The effect of CEOs turnover on the corporate sustainability performance of French firms. *Journal of Business Ethics* 150(4):1049–1069.

Boone T, Ganeshan R, Hicks RL, Sanders NR (2018) Can Google Trends improve your sales forecast? *Production and Operations Management* 27(10):1770–1774.

Bousquet O, Boucheron S, Lugosi G (2004) Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures* 3176:169–207.

Boxenbaum E, Jonsson S (2017) Isomorphism, diffusion and decoupling: Concept evolution and theoretical challenges. *The Sage handbook of organizational institutionalism* 2:79–104.

Brownlee J (2014) Machine learning mastery. URL http://machinelearningmastery.com/discover-feature-engineering-howtoengineer-features-and-how-to-getgood-at-it.

Cachia M, Millward L (2011) The telephone medium and semi-structured interviews: A complementary fit. *Qualitative Research in Organizations and Management: An International Journal* 6(3):265–277.

Calvano L (2008) Multinational corporations and local communities: A critical analysis of conflict. *Journal of Business Ethics* 82(4):793–805.

Chaudhari S, Shevade S (2012) Learning from positive and unlabelled examples using maximum margin clustering. *Proceedings of the 19th international conference on neural information processing-Volume Part III*, 465–473 (Springer-Verlag).

Chen TP (2018) China smites smog with an iron fist. *The Wall Street Journal*. URL https://www.wsj.com/articles/china-smites-smog-with-an-iron-fist-1516185003.

Choi TM, Wallace SW, Wang Y (2018) Big data analytics in operations management. *Production and Operations Management* 27(10):1868–1883.

Chow GC (1993) Capital formation and economic growth in China. *The Quarterly Journal of Economics* 108(3):809–842.

Corbett CJ (2018) How sustainable is big data? *Production and Operations Management* 27(9):1685–1695.

Cormack GV, Hidalgo JMG, Sánz EP (2007) Feature engineering for mobile (SMS) spam filtering. *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, 871–872.

Cui R, Gallino S, Moreno A, Zhang DJ (2018) The operational value of social media information. *Production and Operations Management* 27(10):1749–1769.

Da Silva RFB, Batistella M, Moran EF (2017) Socioeconomic changes and environmental policies as dimensions of regional land transitions in the Atlantic Forest, Brazil. *Environmental Science & Policy* 74:14–22.

Danigelis A (2018) Global emission monitoring systems market to reach $4.44 billion by 2025, report says. URL https://www.environmentalleader.com/2018/06/emission-monitoring-systems-market.

Deepa R, Chezian RM (2014) A study on data cleansing and classification algorithms for large dataset systems. *International Journal of Research in Advent Technology* 2(9):12–22.

Deka GC (2016) Big data predictive and prescriptive analytics. *Big data: Concepts, methodologies, tools, and applications*, 30–55 (IGI Publishing, Herhely, USA).

Delmas M, Keller A (2005) Free riding in voluntary environmental programs: The case of the US EPA WasteWise program. *Policy Sciences* 38(1):91–106.

Dey D, Kumar S (2010) Reassessing data quality for information products. *Management Science* 56(12):2316–2322.

Dey D, Kumar S (2013) Data quality of query results with generalized selection conditions. *Operations Research* 61(1):17–31.

Dubinsky Z (2019) Oilsands CO2 emissions may be far higher than companies report, scientist says. *CBC News.* URL https://www.cbc.ca/news/technology/oilsands-carbon-emissions-study-1.5106809.

Ebenstein A (2012) The consequences of industrialization: Evidence from water pollution and digestive cancers in China. *Review of Economics and Statistics* 94(1):186–201.

ECCC (2020) Canada-United States air quality agreement progress report 2012. *Environment and Climate Change Canada.* URL https://ec.gc.ca/air/default.asp?lang=En&n=8ABC14B4-1&xml=8ABC14B4-ED53-4737-AD51-528F8DBA2B4C&offset=2&toc=hide#s2_2_3.

ECR (2019) Legal enforcement-statistics on environmental infringements. *European Commission Report.* URL http://ec.europa.eu/environment/legal/law/statistics.htm.

Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. *Proceedings of the 14th ACM SIGKDD international conference on Kkowledge discovery and data mining*, 213–220 (ACM).

Elkington J (1998) Partnerships from cannibals with forks: The triple bottom line of 21st-century business. *Environmental Quality Management* 8(1):37–51.

EPA (2016) Control chart methodology for detecting under-reported emissions. *Environmental Protection Agency.* URL https://www.epa.gov/sites/production/files/2018-10/documents/control-chart_method_draft_12-13-16_accessible.pdf.

EPA (2019) 2017 major criminal cases. *Environmental Protection Agency.* URL https://www.epa.gov/enforcement/2017-major-criminal-cases.

EPA (2020) EMC: Continuous emission monitoring systems information and guidelines. *Environmental Protection Agency.* URL https://www.epa.gov/emc/emc-continuous-emission-monitoring-systems.

Fang M, Chan CK, Yao X (2009) Managing air quality in a rapidly developing nation: China. *Atmospheric Environment* 43(1):79–86.

Foster K, Penninti P, Shang J, Kekre S, Hegde GG, Venkat A (2018) Leveraging big data to balance new key performance indicators in emergency physician management networks. *Production and Operations Management* 27(10):1795–1815.

Friedman J, Hastie T, Tibshirani R (2001) *The elements of statistical learning*, volume 1 (Springer series in statistics New York).

Friedman J, Hastie T, Tibshirani R, et al. (2000) Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics* 28(2):337–407.

Fung GPC, Yu JX, Lu H, Yu PS (2005) Text classification without negative examples revisit. *IEEE Transactions on Knowledge and Data Engineering* 18(1):6–20.

Gaber T, Awad A, Ali A, et al. (2020) Feature selection method based on chaotic maps and butterfly optimization algorithm. *International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, 159–169.

Gale A (2018) Nissan admits emissions-test data was falsified. *The Wall Street Journal*. URL https://www.wsj.com/articles/nissan-admits-emission-test-data-was-falsified-1531139749.

Garla VN, Brandt C (2012) Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics* 45(5):992–998.

Goebel P, Reuter C, Pibernik R, Sichtmann C, Bals L (2018) Purchasing managers' willingness to pay for attributes that constitute sustainability. *Journal of Operations Management* 62:44–58.

Gray WB, Shimshack JP (2011) The effectiveness of environmental monitoring and enforcement: A review of the empirical evidence. *Review of Environmental Economics and Policy* 5(1):3–24.

Guha S, Kumar S (2018) Emergence of big data research in operations management, information systems, and healthcare: Past contributions and future roadmap. *Production and Operations Management* 27(9):1724–1735.

Gupta R (2019) Implementation of this efficient pollution monitoring system caught in delays. URL https://www.downtoearth.org.in/news/air/implementation-of-this-efficient-pollution-monitoring-system-caught-in-delays-66746.

He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9):1263–1284.

Hermanns H, Biewer S, D'Argenio PR, Köhl MA (2018) Verification, testing, and runtime monitoring of automotive exhaust emissions. *EPiC Series in Computing* 57:1–17.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8):1735–1780.

Holz CA (2008) Chinas economic growth 1978–2025: What we know today about Chinas economic growth tomorrow. *World Development* 36(10):1665–1691.

Huber PJ (2004) *Robust statistics*, volume 523 (John Wiley & Sons, New York).

Hussain N, Rigoni U, Orij RP (2018) Corporate governance and sustainability performance: Analysis of triple bottom line performance. *Journal of Business Ethics* 149(2):411–432.

Jagadish H, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C (2014) Big data and its technical challenges. *Communications of the ACM* 57(7):86–94.

Jaskie K, Spanias A (2019) Positive and unlabeled learning algorithms and applications: A survey. *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–8 (IEEE).

Ji D, Li L, Wang Y, Zhang J, Cheng M, Sun Y, Liu Z, Wang L, Tang G, Hu B, et al. (2014) The heaviest particulate air-pollution episodes occurred in northern China in January, 2013: Insights gained from observation. *Atmospheric Environment* 92:546–556.

Joglekar NR, Davies J, Anderson EG (2016) The role of industry studies and public policies in production and operations management. *Production and Operations Management* 25(12):1977–2001.

Karplus VJ, Zhang S, Almond D (2018) Quantifying coal power plant responses to tighter SO2 emissions standards in China. *Proceedings of the National Academy of Sciences* 115(27):7004–7009.

Karpoff JM, Lott JR Jr, Wehrly EW (2005) The reputational penalties for environmental violations: Empirical evidence. *The Journal of Law and Economics* 48(2):653–675.

Keller S, Korkmaz G, Orr M, Schroeder A, Shipp S (2017) The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application* 4:85–108.

Kepner J, Gadepally V, Michaleas P, Schear N, Varia M, Yerukhimovich A, Cunningham RK (2014) Computing on masked data: A high performance method for improving big data veracity. *2014 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–6 (IEEE).

Kitchin R (2014) Big data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1):1–12.

Kumar N, Venugopal D, Qiu L, Kumar S (2018) Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems* 35(1):350–380.

Kumar N, Venugopal D, Qiu L, Kumar S (2019) Detecting anomalous online reviewers: An unsupervised approach using mixture models. *Journal of Management Information Systems* 36(4):1313–1346.

Kumar R, Chadrasekaran R (2011) Attribute correction-data cleaning using association rule and clustering methods. *Intl. Jrnl. of Data Mining & Knowledge Management Process* 1(2):22–32.

Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. *Proceedings of the VLDB Endowment* 5(12):2032–2033.

Lau RYK, Zhang W, Xu W (2018) Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Production and Operations Management* 27(10):1775–1794.

Lecué G, Lerasle M, et al. (2020) Robust machine learning by median-of-means: theory and practice. *Annals of Statistics* 48(2):906–931.

Lee G, Xiao X (2020) Voluntary engagement in environmental projects: Evidence from environmental violators. *Journal of Business Ethics* 164:325–348.

Lee HL, Tang CS (2018) Socially and environmentally responsible value chain innovations: New operations management research opportunities. *Management Science* 64(3):983–996.

Lee WS, Liu B (2003) Learning with positive and unlabeled examples using weighted logistic regression. *Proceedings of the 20th international conference on machine learning*, 448–455 (AAAI Press).

Li F, Liu Y, Lü J, Liang L, Harmer P (2015) Ambient air pollution in China poses a multifaceted health threat to outdoor physical activity. *J Epidemiol Community Health* 69(3):201–204.

Li M, Lin Y, Huang S, Crossland C (2016) The use of sparse inverse covariance estimation for relationship detection and hypothesis generation in strategic management. *Strategic Management Journal* 37(1):86–97.

Li M, Wu Y, He Y, Huang S, Nair A (2020) Sparse inverse covariance estimation: a data mining technique to unravel holistic patterns among business practices in firms. *Decision Sciences* 51(4):1046–1073.

Li X, Liu B (2003) Learning to classify texts using positive and unlabeled data. *Proceedings of the 18th international joint conference on artificial intelligence*, 587–592 (Morgan Kaufmann Publishers Inc.).

Lin Y (2004) A note on margin-based loss functions in classification. *Statistics and Probability Letters* 68(1):73–82.

Liu L, Peng T (2014) Clustering-based method for positive and unlabeled text categorization enhanced by improved TFIDF. *Journal of Information Science & Engineering* 30(5):1463–1481.

Lu F, Bai Q (2010) Semi-supervised text categorization with only a few positive and unlabeled documents. *2010 3rd International Conference on Biomedical Engineering and Informatics*, volume 7, 3075–3079 (IEEE).

Marquis C, Qian C (2014) Corporate social responsibility reporting in China: Symbol or substance? *Organization Science* 25(1):127–148.

Mendelssohn IA, Andersen GL, Baltz DM, Caffey RH, Carman KR, Fleeger JW, Joye SB, Lin Q, Maltby E, Overton EB, et al. (2012) Oil impacts on coastal wetlands: Implications for the Mississippi River Delta ecosystem after the Deepwater Horizon oil spill. *BioScience* 62(6):562–574.

Meyer JW, Rowan B (1977) Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology* 83(2):340–363.

Mishra D, Gunasekaran A, Papadopoulos T, Childe SJ (2018) Big data and supply chain management: A review and bibliometric analysis. *Annals of Operations Research* 270(1-2):313–336.

Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62:22–31.

Murray A, Skene K, Haynes K (2017) The circular economy: An interdisciplinary exploration of the concept and application in a global context. *Journal of Business Ethics* 140(3):369–380.

Nonet G, Kassel K, Meijs L (2016) Understanding responsible management: Emerging themes and variations from European business school programs. *Journal of Business Ethics* 139(4):717–736.

Oppel Jr RA (2000) Environmental tests falsified, US says. *New York Times A* 14, URL https://www.nytimes.com/2000/09/22/us/environmental-tests-falsified-us-says.html.

Ozdemir S, Susarla D (2018) *Feature engineering made easy: Identify unique features from your dataset in order to build powerful machine learning systems* (Packt Publishing Ltd).

Pan L, Wang Zk, Wang Zx (2005) Present status and countermeasure suggestion for thermal power plants CEMS in China. *Research of Environmental Sciences* 18(4):42–45.

Plambeck E, Lee HL, Yatsko P (2012) Improving environmental performance in your Chinese supply chain. *MIT Sloan Management Review* 53(2):43.

Porteous AH, Rammohan SV, Lee HL (2015) Carrots or sticks? Improving social and environmental compliance at suppliers through incentives and penalties. *Production and Operations Management* 24(9):1402–1413.

Rhodes C (2016) Democratic business ethics: Volkswagens emissions scandal and the disruption of corporate sovereignty. *Organization Studies* 37(10):1501–1518.

Roque PG, Severo EA, Dorion ECH, Roque EdS, de Guimarães JCF (2018) The dilemma of environmental sustainability in a developing country: Environmental crimes in southern Brazil. *Business Strategy & Development* 1(1):43–52.

Rozados IV, Tjahjono B (2014) Big data analytics in supply chain management: Trends and related research. *6th International Conference on Operations and Supply Chain Management, Bali*, volume 1, 2013–2014.

Rubin V, Lukoianova T (2013) Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online* 24(1):4.

Sanders NR, Ganeshan R (2018) Big data in supply chain management. *Production and Operations Management* 27(10):1745–1748.

Shmueli G, Koppius O (2011) Predictive analytics in information systems research. *MIS Quarterly* 35(3):553–572.

Shmueli G, Yahav I (2018) The forest or the trees? Tackling Simpson's Paradox with classification trees. *Production and Operations Management* 27(4):696–716.

Siano A, Vollero A, Conte F, Amabile S (2017) More than words: Expanding the taxonomy of greenwashing after the Volkswagen scandal. *Journal of Business Research* 71:27–37.

Singhal K, Feng Q, Ganeshan R, Sanders NR, Shanthikumar JG (2018) Introduction to the special issue on perspectives on big data. *Production and Operations Management* 27(9):1639–1641.

Singhal K, Singhal J (2019) Technology and manufacturing in China before the industrial revolution and glimpses of the future. *Production and Operations Management* 28(3):505–515.

Sodhi MS (2015) Conceptualizing social responsibility in operations via stakeholder resource-based view. *Production and Operations Management* 24(9):1375–1389.

Strong DM, Lee YW, Wang RY (1997) Data quality in context. *Communications of the ACM* 40(5):103–110.

Swaminathan JM (2018) Big data analytics for rapid, impactful, sustained, and efficient (RISE) humanitarian operations. *Production and Operations Management* 27(9):1696–1700.

Tang CS (2018) Socially responsible supply chains in emerging markets: Some research opportunities. *Journal of Operations Management* 57(1):1–10.

Tang L, Qu J, Mi Z, Bo X, Chang X, Anadon LD, Wang S, Xue X, Li S, Wang X, et al. (2019) Substantial emission reductions from Chinese power plants after the introduction of ultra-low emissions standards. *Nature Energy* 4(11):929–938.

Wang G, Gunasekaran A, Ngai EW, Papadopoulos T (2016) Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics* 176:98–110.

Wang RY, Strong DM (1996) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12(4):5–33.

Xia L, Guo T, Qin J, Yue X, Zhu N (2018) Carbon emission reduction and pricing policies of a supply chain considering reciprocal preferences in cap-and-trade system. *Annals of Operations Research* 268(1):149–175.

Xu X, Zeng S, Tam CM (2012) Stock markets reaction to disclosure of environmental violations: Evidence from China. *Journal of Business Ethics* 107(2):227–237.

Xu X, Zeng S, Zou H, Shi JJ (2016) The impact of corporate environmental violation on shareholders' wealth: A perspective taken from media coverage. *Business Strategy and the Environment* 25(2):73–91.

Yin S, Kaynak O (2015) Big data for modern industry: Challenges and trends. *Proceedings of the IEEE* 103(2):143–146.

Zhang D, Lee WS (2005) A simple probabilistic approach to learning from positive and unlabeled examples. *Proceedings of the 5th annual UK workshop on computational intelligence (UKCI)*, 83–87.

Zhang X, Schreifels J (2011) Continuous emission monitoring systems at power plants in China: Improving SO2 emission measurement. *Energy Policy* 39(11):7432–7438.

Zou H, Zeng R, Zeng S, Shi JJ (2015) How do environmental violation events harm corporate reputation? *Business Strategy and the Environment* 24(8):836–854.