APPLICABILITY OF DATA DRIVEN METHODS FOR ASSESSING COMPLIANCE OF WASTEWATER TREATMENT PLANTS SELF-REPORTED DATASETS

By

Pouyan Hatami Bahman Beiglou

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biosystems Engineering- Master of Science

2016

ABSTRACT

APPLICABILITY OF DATA DRIVEN METHODS FOR ASSESSING COMPLIANCE OF WASTEWATER TREATMENT PLANTS SELF-REPORTED DATASETS

By

Pouyan Hatami Bahman Beiglou

The primary source of compliance information in water quality monitoring is self-reported data. Despite the heavy reliance on self-reported data in United States environmental regulation, the U.S. General Accounting Office has expressed concerns regarding the potential for fraud in environmental self-reports. Furthermore, recent research indicates that the methods used by state enforcement are unlikely to detect fraud. Therefore, the need for data-driven methods to support regulatory enforcement is an important area of research. In this thesis, we evaluated the applicability of data-driven methods for assessing compliance of wastewater treatment plants (WWTP) self-reported datasets based on a description of the variability in these data streams. For this purpose, first a literature review was conducted (1) to determine the goals of the Clean Water Act programs; (2) identify limitations of current monitoring efforts and data gaps in the understanding of the sources of variability in WWTPs data; and (3) to identify appropriate predictive analytical methods to address the problems. Second, the applicability of Benford's Law as a method for uncovering irregularities in the distribution of first and second digits in a sample dataset was tested and its effectiveness was discussed. Finally, the use of other promising approaches, which may be capable of finding mishandling in wastewater treatment plants are presented with preliminary data.

ACKNOWLEDGEMENTS

I would like to thank my advisors Dr. Jade Mitchell and Dr. Amir Pouyan Nejadhashemi for all their support, guidance and advice. Without them, I would not have been able to write this thesis. I would also like to thank my committee, Dr. Timothy Harrigan and Dr. Carole Gibbs for their guidance in completing this degree.

TABLE OF CONTENTS

LIST OF TABLES

KEY TO ABBREVIATIONS

CI - Confidence Interval

CWA - Clean Water Act

EPA - Environmental Protection Agency

GAO - General Accounting Office

GPD - Gallons Per Day

MGD - Million Gallons per Day

NPDES - National Pollutant Discharge Elimination System

PCS - Permit Compliance System

TMDL - Total Maximum Daily Load

WWTP - Wastewater Treatment Plants

WERF - Water Environmental Research Foundation

1 INTRODUCTION

Amendments to the 1972 Clean Water Act, the foundation of surface water quality protection in the United States establish the supreme goal of this legislation “to restore and maintain chemical, physical and biological integrity of the Nation’s water.” As such, Section 402 of the Clean Water Act established the National Pollutant Discharge Elimination System (NPDES), which regulates the level of pollutant discharges from point sources into the waters of the U.S. through permitting programs. Permits are defined based on Total Maximum Daily Load (TMDL), which determines the maximum amount of a pollutant that can occur in a waterbody. All dischargers are required to receive a permit before discharging their effluents into the surface waters.

Monitoring the compliance of dischargers is required in every state, but current strategies rely heavily on inspections, which occurs yearly, biannually or as infrequently as once every two years or once every five years for major dischargers and minor dischargers, respectively. Therefore, environmental monitoring and enforcement relies extensively on regulated entities to self-report pollution discharges. The only compliance evaluation of the self-reported datasets is through visual monitoring whereas studies have reported that compliance evaluation of self-reported data needs a more in-depth analysis (Shimshack & Ward, 2005).

Due to the lack of a robust platform to assess the integrity of self-reported data, human and ecological health are potentially endangered by risks associated with underreported discharges. Therefore, the need for data-driven methods to support regulatory enforcement is an important compliance evaluation area of research. Data-driven[1] methods can provide an effective approach

[1] Data-driven methods are the methods that are based on data rather than intuition or personal experience

and to ensure achieving water quality safety by finding the underlying patterns and relationships within the data, cost lower than inspections for regulatory enforcement. Furthermore, data-driven methods are more simple to implement compared to the conventional monitoring methods, and may allow for more timely regulatory enforcement.

Therefore, the specific objectives of this thesis are as follows:

1) To review the current monitoring processes and identify data gaps and limitations;
2) To assess and test the applicability of a simple data-driven approach using predictable wastewater treatment plants (WWTPs) data;
3) To discuss the effectiveness of the approach;
4) To discuss other promising data-driven approaches.

2 WATER QUALITY MONITORING FRAMEWORK

2.1 WATER RESOURCES MANAGEMENT

One of the most important aspects of water resources management is to maintain water quality. Unfortunately, industrialization and urbanization have led to many water quality issues through point sources and non-point sources (Wang, 2001). Because surface water pollution is a major water quality problem in the United States, regulated and unregulated programs have been developed by the United States Environmental Protection Agency (USEPA) to control the amount of pollution into the surface water (Parry, 1998). One of the primary regulations put in place to control the amount of discharges into surface waters is the Clean Water Act.

2.1.1 Clean Water Act

The Clean Water Act, initially executed in 1948 as the Federal Water Pollution Control Act, was the key element of surface water quality protection in the United States, but was significantly reorganized and developed in 1972 (EPA, 2002). Prior to 1987, it was exclusively directed at point source pollution, but amendments to the law in that year included nonpoint source pollution measurements. Point source pollution has caused more than 50% of the country's water quality problems (Copeland, 2010). Non-point source pollution has been identified as the likely the most significant source of water pollution (Puckett, 1995). Due to the already extensive regulations, point source pollution has generally has received less attention as an ongoing threat, but it may be more harmful than anticipated (Andreen, 2004). Specifically, point source discharges hidden by fraudulent self-reported data represent another potential source of water quality impairment.

Although the Clean Water Act was specifically established to protect water quality or to restore degraded water, many years after its enactment, the nation's aquatic ecosystem is still stressed (Doremus & Dan Tarlock, 2012). Title VI of the Clean Water Act explains that all industrial and municipal dischargers are prohibited from violating water quality standards, and that states are responsible for ensuring all regulatory requirements are met (Copeland, 2014). Water quality standard are defined through a program called TMDL.

2.1.2 Total Maximum Daily Load (TMDL)

The TMDL program rose as an establishment for the country's endeavors to meet state surface water quality standards. Total maximum daily load of a pollutant is the amount of discharge that complies with a water quality standard; the "TMDL process" refers to the arrangement to create and perform the TMDL (Tjeerdema, 2007). A TMDL is the measure of pollutant loads apportioned from point sources and nonpoint sources in addition to a margin of safety for probable unknown and seasonal changes in water quality (Miller-mcclellan, Shanholtz, & Miller-mcclellan, 2003). The goal of the TMDL is to guarantee that the waterbody will have the capacity to meet water quality standards for all seasonal variations.

Generally, TMDL point and nonpoint loads are evaluated using computer modeling. Although monitoring is the most desirable method to calculate TMDL loads, its utilization is limited because of the high cost and the large variability in spatial and temporal components of ecosystems that would necessitate a prohibitive amount of samples to fully characterize the water quality of a waterbody (Muñoz-Carpena et al., 2006). The margin of safety is incorporated to represent uncertainties connected with the improvement of the TMDL and is added to increase water quality protection (Zhang & Yu, 2004).

2.1.2.1 TMDL Deficiencies

Despite the importance of the TMDL program, its advancement is still a challenging task, particularly when there is no technical guidance to deliver help in executing uncertainty analysis

(Shirmohammadi, Chaubey, & Harmel, 2006). Theoretically, due to the regulatory and strategic decision-making processes, the uncertainty analysis always has some arbitrariness so as a long-term consequence, that can affect the success of the TMDL program (USGS, 2008). Furthermore, due to tight timetables and constrained financial resources for TMDL improvement, typically the margin of safety has been selected by subjective decisions without clear consideration of uncertainty sources and estimation their direct influences on total uncertainty in the TMDL calculations (USGS, 2008).

A national study supported by the Water Environmental Research Foundation (WERF) revealed that among 172 TMDLs, 12 TMDLs had no margin of safety estimates at all; 119 of the remaining TMDLs used the subjective EPA simple explicit margin of safety technique; 40 of them had conservative assumptions, and only one TMDL unambiguously calculated the uncertainty over a analogous research study and redirected this uncertainty into Margin of safety (Dilks & Freedman, 2004).

While TMDLs are being established, discharge permits may be issued to dischargers through the National Pollutant Discharge Elimination System (NPDES) program.

2.1.3 National Pollutant Discharge Elimination System (NPDES)

Under section 402 of the Clean Water Act, the National Pollutant Discharge Elimination System requires the acquisition of a permit[2] for all facilities discharging wastewater into the surface water of the nation (EPA, 2004). NPDES has played an important role in protecting and restoring water quality in the United States by regulating and limiting direct discharge into the

[2] There are two types of individual and general permits which a facility can request where each permit type is used under different conditions and embrace different permit issuance procedures (Boyd, 2003). An individual permit is a permit that is issued for an individual facility based on the its information, such as previous permit requirements, discharge monitoring reports, technology, water quality standards and total maximum daily loads, whereas a general permit covers multiple facilities, which can be considered in a specific group of dischargers (Gaba, 2007).

surface water (Houck, 2002). While in 1972, only about 30% of the United States surface water were considered to be healthy, this was increased to approximately two thirds by 2001 (Birkeland, 2001).

A facility owner or operator has to apply for an NPDES permit through EPA or a state permitting authority. Then the permit writer defines the proper permit terms and conditions through evaluating facility-specific information (Gaba, 2007).

As of 2010, more than 65,000 industrial and municipal dischargers must attain NPDES permit from the EPA or qualified states (Copeland, 2010).

2.1.3.1 NPDES Deficiencies

The main performance gap of NPDES is outdated permits (Rechtschaffen & Markell, 2003). Facilities can use their outdated permits as long as the request for permit renewal is under process. As of 2003, 15% of major facilities and one third of minor facilities were using outdated permits.

2.2 ENVIRONMENTAL MONITORING PRACTICE AND IMPLEMENTATION

Monitoring is needed for policy makers to plan, develop and assess environmental rules. A monitoring program is designed to ensure quality and accessibility of data and cost-effectiveness evaluation of water quality protection programs (Lovett, Burns, & Driscoll, 2007). The NPDES monitoring and reporting conditions section defines detailed requirements for location and frequency of monitoring, sample collection techniques, analytical methods, reporting and recordkeeping (EPA, 2010a).

Section 308 of the Clean Water Act, which authorizes monitoring of facilities to ensure that water quality standards are met, provides two types of monitoring (EPA, 2004):

1- Self-monitoring, where the facility must monitor wastewater components alone;

2- Monitoring by the EPA or the state, which consists of two processes:

(i) Evaluation of facility self-monitoring; and

(ii) Direct monitoring activities.

Regardless of what type of monitoring (inspections, etc.), the frequency of monitoring the discharge is related to several factors. These factors include design capacity of the treatment facility, compliance history, treatment method used, cost of monitoring relative to permittee's capability, discharge location, types of pollutants, frequency of discharge, and sum of monthly samples used in developing effluent limitations (EPA, 2010). For example, a highly variable discharge should be monitored more frequently than a discharged water quality parameter, which is more consistent over time. Data collected from tracking plant-level self-reported emissions and on-site inspections along with permitted effluent limitations and enforcement action are saved in the EPA's permit compliance system (PCS).

Monitoring programs have showed success since they were initiated, and empirical studies have found that they positively influence compliance and levels of pollutant discharge. Magat and Viscusi (1990) evaluated the impact of monitoring on water quality of 77 pulp and paper mills between 1982 to 1985. They found that while their overall compliance rate was about 75%, not inspecting the facilities in the previous quarter could double the possibility of noncompliance. Also, they assessed the impact of monitoring on the amount of discharge pollution by facilities, which was about 20% decrease for each inspection. Earnhart (2004) and Glicksman & Earnhart (2007) examined conventional water pollution discharge for forty Kansas wastewater treatment plants and hundred chemical facilities respectively. Both studies found that monitoring programs along with monetary fines steadily decreased relative discharges. Shimshack & Ward (2005) assessed the compliance of 217 pulp and papers facilities between

1988 to 1996 after penalizing and regulatory action. Application of additional fines caused two-thirds of water pollution violation percentage drop in the year following the actions.

2.2.1 Deficiencies in Environmental Monitoring Practice and Implementation

Although there have been endeavors to monitor the compliance of water pollution dischargers, it is not a flawless system and noncompliance is threatening waterbodies. Most monitoring practices perform as expected and there are still controversial topics among practitioners, regulators and researchers (Harmancioglu, Fistikoglu, Ozkul, Singh, & Alpaslan, 1999). There are reports about permits violation (GAO, 1983), in which 82% of dischargers violated their permit at least once. Additionally, 24% were in a substantial noncompliance status with their discharge permit. The U.S. General Accounting Office (GAO) report of fiscal years 1992-1994 also declared that one in six major facilities was significantly in noncompliance with its allocated discharge permit and that the actual number could be twice as high. A nationwide compliance analysis of major facilities performed by the EPA disclosed that 25% were significantly in noncompliance with their discharge permits at any given time (Rechtschaffen & Markell, 2003).

Severe deficiencies of the Clean Water Act’s monitoring program have been pointed out in many studies (Glicksman & Earnhart, 2007; Rechtschaffen & Markell, 2003; Shimshack & Ward, 2005; Magat and Viscusi, 1990) such as failure to perform inspections, failure to implement the proper actions; and failure to have effective penalties. Gray & Shimshack (2011) declared another weakness of the monitoring programs due to significant variability across time and state authorities such as different inspection frequencies and noncompliance fines. They discussed that cross-state variability in facilities composition leads to defining non-practical federal enforcement guidelines and monitoring strategies. However, few studies have addressed this variability within or across states.

2.2.2 Discharge Monitoring Report Review

Self-monitoring reports are considered the primary source of information for permittee compliance evaluation (Shimshack & Ward, 2005). Monitoring programs require the permit holders to self-report their water pollution discharge routinely and report the analytical results to the permitting authority with the essential information to assess discharge characteristics and compliance status (EPA, 2010b). Periodic self-reporting creates a continuous record of the permittee's compliance status, which can help to detect violations as well as providing a source of information to support any necessary enforcement action (NYSDEC, 2012).

Facilities report their self-monitoring to the permitting authorities and the permitting authorities are responsible to transfer the facilities report to EPA headquarters either electronically or manually. Subsequently, the reports are entered into the EPA electronic database to be reviewed for any permit noncompliance (EPA, 2006).

2.2.2.1 Deficiencies in Self-Monitoring Report System

Despite the heavy reliance on self-monitoring data, the GAO has expressed concerns regarding the potential for fraud in environmental self-reports (GAO, 1993). Strategic misreport can lead to inaccurate compliance evaluation and consequently inaccurate estimation of the discharge pollutions putting human and ecology in danger. Also, averaged reported values over periods of time (i.e. weekly, monthly and quarterly) may cause inaccurate estimation of real time discharges and be another reason for potential environmental risks.

Because self-reported violations are treated with administrative penalties and the outcome of strategic falsification of reports can lead to criminal prosecution of employees and managers, generally self-monitoring reports are considered to be truthful (Gray & Shimshack, 2011). Kaplow & Shavell (1991) discussed that dischargers can be encouraged to report their own violations without materially affecting their motivation to refrain from violation. Shimshack & Ward (2005) discussed that veracity assessment of a facility's self-reporting records are carried out through visual monitoring whereas compliance assessment mandates a more in-depth analysis. The authors posit that though on-site state or federal EPA inspections are another method to ensure the accuracy of self-report records and verify the maintenance and operation of facilities, permittees are intended to report truthfully in the presence of regulatory inspection. The researcher proposed a secret and random inspection of the facilities as an ideal solution (Gray & Shimshack, 2011).

2.2.2.2 Malfunction in Wastewater Treatment Plants, Reasons of Inaccurate Self-Reporting

The existence of problems in wastewater treatment plants may cause violations in the reporting of the actual effluent discharges. More than two third of the wastewater treatment plants in the United States had serious gaps in their measurements and more than one thirds of all sewage systems have been in noncompliance with environmental laws (Duhigg, 2009). Flajsig (1999) studied common problems in wastewater treatment plants. The malfunctions may be due to poor quality of planning and designing data, lack of experience of plants operators, poor maintenance of the plants due financial problems. Also, Freeman (1990) presented maintenance and equipment deficiencies, and treatment plants overloading as the reasons for violation in operation of wastewater treatment plants.

Because of all those problems, there are strict regulations for wastewater treatment plants to employ new technologies and optimize the current ones but applicability and financial aspects of the change should be considered (Jury & Vaux, 2005).

2.2.3 On-Site Compliance Evaluation

On-site monitoring ranges from quick inspections for few hours to a very comprehensive inspection, which can last to up a month or more (Gray & Shimshack, 2011). Current EPA compliance monitoring strategies require major and minor facilities to receive a comprehensive inspection at least once biannually and every five years, respectively. However, regulators are not limited to those timelines and the frequency of compliance assessment can be increased (Tsakiris & Alexakis, 2012).

2.2.3.1 Deficiencies in On-Site Compliance Assessment

On-site monitoring has several deficiencies, which may lead to facility non-compliance. According to the EPA's Office of Inspector General, due to lack of knowledge about the actual size of the facilities and how their sizes have changed over time, frequency and comprehension of the inspection strategies needs to be matched with their development rates (EPA, 2005). Regulatory agencies are also suffering from lack of budget and local economic circumstances (Deily & Gray, 1991), pressures from local interest groups (Peltzman, 1976), and political burdens (Kleit, Pierce, & Hill, 1998), which can result in circumstances where performing an appropriate and more frequent inspection is unachievable.

Storey et al. (2011) studied inspection methods as another area of discussion for deficiencies in on-site compliance assessment. They argued that although, recently there have been promising technological developments in biological monitors and micro sensors to monitor water quality and contaminant detection, large-scale implementation will still take years to achieve. While employing advanced monitoring technologies comes at a high cost and is non-compatible with current treatment operations, they need to evolve to meet many operational limitations.

Earnhart (2010) analyzed the effects of permitted discharge limits on strategy of inspection in the municipal wastewater treatment plants both empirically and theoretically. He found that when relative discharge of a facility increases, there is a greater probability of noncompliance for the facility so in result the likelihood of the inspection by the agencies in the preceding month rises. In turn, Shimshack & Ward (2005) discussed that theoretically a plant may quickly reduce its effluents to the standard levels when there is the possibility that regulatory inspectors be present.

2.2.3.2 Deficiencies in Use of Sampling Studies in Defining Monitoring Strategies

Before monitoring it should be considered that water quality monitoring is a highly complex process. This complexity arises from uncertainty in the nature of water quality and uncertainty in defining the specific purpose of monitoring (Harmancioglu et al., 1999). Uncertainties in the nature of water quality are due to natural hydrologic cycle and human-made influences (Sanders, 1983). Spatial and seasonal variations of water quality are extremely relevant to land use pattern and influences from watershed runoff discharge (Caccia & Boyer, 2005; Ming-kui, Li-ping, & Zhen-li, 2007). Horowitz (2013) discussed the result of those uncertainties as biased sampling, processing and analytical methods, which generate nonrepresentative data. He argued that the assumptions of the existing programs such as calendar-based sampling and stationarity are not defensible anymore, some monitoring programs may need to be redesigned, some sampling and analytical methods need to be updated (e.g. sampling locations and frequency) and statistical models which do not consider dynamic characteristics of hydrologic interrelationships which may require recalibration.

2.2.4 Use of Remote Sensing in Monitoring

Remote sensing can be defined as the “a science and art of obtaining information about an object, area, or phenomena through the analysis of data acquired by device that is not in contact with the object, area, or phenomena under investigation” (Lillesand, Kiefer, & Chipman, 2014). There are various studies which discussed about the usefulness of remote sensing as a method of monitoring the water quality (Giardino, Brando, .& Dekker, 2007; Ritchie & Cooper, 1988; Schalles, Gitelson, Yacobi, & Kroenke, 1998). Use of remote sensing in water quality monitoring has been conducted since early 1970’s.

There are many studies in which each one describes a different satellite sensor (Alparslan, Aydöner, Tufekci, & Tüfekci, 2007; Giardino et al., 2007; He, Chen, Liu, & Chen, 2008; Maillard & Santos, 2008). In each, attributes of water and pollutants are crucial to water quality monitoring. Received signal has a spectral characteristic, which is a function of hydrological, biological and chemical features of water (Seyhan & Dekker, 1986). For example suspended solids cause an increase in radiance emergent from surface waters near infrared quantity of the electromagnetic spectrum (Ritchie & Cooper, 1988).

2.2.4.1 Deficiencies in Use of Remote Sensing in Monitoring

It is very usual to have systematic errors in use of remote sensing as a tool to monitor water quality due to reflectance terminology does not convey physical standards entirely (Schaepman-Strub, Schaepman, Painter, Dangel, & Martonchik, 2006) so an analysis of the systematic errors along with random errors of the data is required. The revision of systematic errors is an essential for the correction of the depth and color data, and depends on the identification of the mathematical model and included parameters (Khoshelham, 2011). Another source of error in remote sensing monitoring may originate from the sensor, measurement settings and characteristics of the object surface (Khoshelham, 2011).

2.3 SUMMARY

Although there have been endeavors to monitor the compliance of water pollution dischargers, still it is not flawless and noncompliance is threatening waterbodies. Most monitoring practices not perform what is expected and there are still controversial topics among practitioners, regulators and researchers. Deficiencies in monitoring programs originate from inappropriate uncertainty analysis in TMDL, using outdated NPDES permits, possible falsification in self-monitoring reports, old treatment and measurement technologies in wastewater treatment plants, and improper inspection frequency and methods. Consequently, incomplete and inaccurate generated data affect the excellence of monitoring program.

A significant amount of environmental regulation in the United States is conducted via self-reports but still the U.S. GAO has expressed concerns regarding the potential for fraud in environmental self-reports (US GAO, 1993), which represents a criminal act under many environmental laws. The most recent EPA Office of Inspector General (OIG) report indicates that efforts to improve procedures to detect and address environmental reporting fraud remain inadequate (1999; 2014). The US EPA delegated regulatory authority for most major federal environmental programs to the states (Situ & Emmons, 2000). The environmental regulatory offices tasked with assessing the level of compliance of permitted entities have limited resources to proactively assess the veracity of the data (Dumas & Devine, 2000).

The potential for fraud in the self-report process is an important issue. Despite the laws in place to maintain water body quality standards, water resources are still threatened (Parry, 1998). Two-thirds of coastal systems, one-third of streams and two-fifths of lakes in the United States are impaired due to nutrient loading (Davidson et al., 2011). Non-point source pollution has been identified as the likely the most significant source of water pollution (Puckett, 1995).

Due to the already extensive regulations, point source pollution has generally has received less attention as an ongoing threat, but it may be more harmful than anticipated (Andreen, 2004). Specifically, point source discharges hidden by fraudulent self-reported data represent another potential source of water quality impairment.

Therefore, it is important to devise strategies to discover and address fraudulent self-reports in the environmental arena, particularly at the state-level. Although understanding of state regulatory and enforcement processes is limited, recent research indicates that state methods (e.g., the on-site inspection process) are unlikely to detect fraud (Rivers, Dempsey, Mitchell, & Gibbs, 2015). New solutions that complement and can be easily integrated into existing state practices are needed (Rivers et al., 2015). This thesis represents a direct effort to build on these suggestions by examining data-driven methods to detect potential fraud.

3 ANALYSIS AND RESULTS OF THE FIRST STUDY

3.1 INTRODUCTION

In this chapter, we evaluate Benford's Law as a method for uncovering fraud in water discharge self-report data. Benford's Law has been broadly applied in many fields (M. Nigrini, 2012) and has been used with some success to detect irregularities in environmental data, such as concentrations of chemical emissions to air (De Marchi & Hamilton, 2006) and air quality monitoring data (Fu, Fang, Villas-Boas, & Judge, 2014). Given the relative ease of using Benford's Law coupled with state resource constraints, this work is an important starting point for evaluating data driven methods to detect fraud in environmental self-reports. This paper represents a direct effort to build on these suggestions by examining data-driven methods to detect potential fraud.

In the 1930s, physicist Frank Benford found a first digit pattern of numbers in certain datasets(Mark J Nigrini & Miller, 2007), which later became known as Benford's Law. The law describes the expected frequency of the numbers between one and nine in which the appearance of lower numbers (i.e. 1, 2, and 3) as first digits are more likely to appear than higher numbers (i.e. 7, 8 or 9) as first digits within a given dataset. The expected frequency of the first digits reported in a dataset follow a logarithmic pattern (Benford, 1938).The formula to describe the discrete probability distribution of the frequency of occurrence of single digit numbers as first digits (Brown, 2005).

The law describes the lower numbers (i.e. 1, 2, and 3) as first digits appear more than higher (i.e. 7, 8 or 9) as within a given dataset. The expected frequency of the first digits reported in a dataset follow a logarithmic pattern (Benford 1938) as equation 1:

$$P(d_1) = \log[1 + \left(\frac{1}{d_1}\right)]; \qquad d_1 \epsilon \{1,2,\dots,9\} \qquad (1)$$

where, P is the expected probability of a number with a first digit equal to d_1 and $P(d_1)$ for all possible first digits is tabulated in Table 1.

Table 1- The expected probability of first significant digit (FSD) predicted by Benford's Law

First Digit	**Probability (%)**
1	30.1
2	17.6
3	12.5
4	9.7
5	7.9
6	6.7
7	5.8
8	5.1
9	4.6

Benford’s Law has also been expanded to describe the frequency for the first two digits as described by Equation 2 (Mark J Nigrini, 2005):

$$P(d_1 d_2) = \log[1 + \left(\frac{1}{d_1 d_2}\right)] \quad d_2 \epsilon \{0,1,2,\dots,9\} \qquad (2)$$

A continuous function to describe the expected probabilities of the appearance of the first two digits is presented in the Table 2.

Table 2- Expected Probability of the First Two Digits Predicted by Benford’s Law

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	4.14	3.78	3.48	3.22	3.00	2.80	2.63	2.48	2.35	2.23
2	2.12	2.02	1.93	1.85	1.77	1.70	1.64	1.58	1.52	1.47
3	1.42	1.38	1.34	1.30	1.26	1.22	1.19	1.16	1.13	1.10
4	1.07	1.05	1.02	1.00	0.98	0.95	0.93	0.91	0.90	0.88
5	0.86	0.84	0.83	0.81	0.80	0.78	0.77	0.76	0.74	0.73
6	0.72	0.71	0.69	0.68	0.67	0.66	0.65	0.64	0.63	0.62
7	0.62	0.61	0.60	0.59	0.58	0.58	0.57	0.56	0.55	0.55
8	0.54	0.53	0.53	0.52	0.51	0.51	0.50	0.50	0.49	0.49
9	0.48	0.47	0.47	0.46	0.46	0.45	0.45	0.45	0.44	0.44

3.1.1 Criteria for Application of Benford’s Law

In general, several standards should be considered before applying Benford’s Law to a given dataset (Dumas & Devine, 2000). If the standards are ignored, correctly labeling a dataset as fraudulent or anomalous cannot be accurately inferred if the dataset does not follow Benford’s Law (Durtschi, Hillison, & Pacini, 2004). One of the most important considerations is the nature of the dataset and the context of the area of focus from which it pertains. A dataset should “describe a single type of phenomena” (Dumas & Devine, 2000)(e.g. wastewater treatment plant discharge parameters). The dataset should not have an inherent minimum and maximum (Durtschi et al., 2004). For example, the hydrogen ion concentration of a solution (pH) varies between 0-14 so a dataset self-reporting pH as a parameter could not be considered in Benford’s Law assessment. A dataset should also be spread across at least one order of magnitude (expressed as a power of 10)(Dumas & Devine, 2000). Some researchers refer to this criterion as “too uniform” (Brown, 2005). Wallace (2002) suggested if the mean of the dataset is greater than the median and the data skewness of the distribution is positive, it is more likely to obey Benford’s Law. Another criterion or standard is that the number of reported values should not be “too small” (Brown, 2005). When a dataset has few reported values, the criterion of the frequency of first digits being spread across at least one order of magnitude will not be satisfied.

Whereas Benford’s Law can be incredibly helpful to assess the veracity of datasets, it has limitations that must be considered before applying the Law.

3.1.2 Previous Applications of Benford’s Law

In accounting and financial data, Carslaw (1988) applied Benford’s Law on earning numbers of New Zealand firms and proved that the datasets reported were biased based on significantly different distributions than Benford’s Law distribution. In fraud detection in accounting numbers, Nigrini (1992, 1994, 1996, 1999, 2003, 2005, 2007, 2011, 2012) is the first researcher who has extensively employed Benford’s Law as a tool to find data irregularities. Benford’s Law has also been used to find anomalies in the results of elections. For example, in 2009 presidential election in Iran, the Benford’s Law analysis of vote counts for four candidates in 366 voting areas showed evidence of fraud (Roukema, 2009). Beyond the social sciences, numbers and digits are often lognormally distributed in nature and as a result, the enormous prevalence of Benford’s Law can be seen in physical sciences including physics, chemistry, astronomy, geology, and biology (Kossovsky, 2015). Sambridge et al. (2010) proved applicability of Benford’s Law to natural science observations such as the earthquake depths, time between earthquakes, rotation frequencies of pulsars, river lengths of Canada, global temperature anomalies, greenhouse gas emissions, global infectious diseases reported by the World Health Organization, the Earth’s geomagnetic field, time between geomagnetic reversal, seismic body P-wave speeds of Earth's mantle, the brightness of gamma rays reaching Earth.

Benford’s Law has also been used extensively in the environmental arena, including self-reported datasets. Docampo et al. (2009) showed the gross datasets of daily pollen counts from three aerobiological stations fit Benford’s Law for the first significant digits. Data was taken from stations located in European cities with different vegetation and climatology. Vries and

Murk (2013) applied Benford’s Law for the first time to ecotoxicological data from the U.S. EPA ECOTOX database, LC50 (calculation of lethal concentration at which 50% of the subject succumb) and NOEC (no observed effect concentration) as a tool to quickly screen large amounts of data for irregularities in order to identify the reliability of the data for risk assessment. The methodology identified deviations from Benford’s Law for large datasets of interpolated NOEC. Brown (2005) assessed Benford’s Law as a potential data screening and authenticity checking tool for several datasets related to the measured concentrations of pollutants in ambient air in the U.K. Analysis of the first digit of the dataset found some datasets having a very close fit to Benford’s Law distribution; however, some fit poorly based on their orders of magnitude[3]. In that study, a plot of the numerical range of datasets versus the sum of the normalized deviation from Benford’s Law illustrated that datasets containing numbers across a minimum of four orders of magnitude had a very close conformity to Benford’s Law, whereas fewer than four orders of magnitude had an exponential reduction in correlation with the law.

De Marchi and Hamilton (2006) compared the first digit distribution of air emissions reported by plants in the Toxic Release Inventory with chemical concentration levels measured by EPA pollution monitors using Benford’s Law analysis. The results suggested two heavily regulated chemicals, lead and nitric acid were not accurately self-reported. In 2014, Fu et al. reported results from using Benford’s Law to track Air Quality Index data for 35 monitors in Beijing. A substantial number of Beijing monitors reported daily air quality that departed significantly from Benford’s Law distribution, while the hourly real-time data fit well. Through the use of principal components ,which had higher traffic volumes and housing prices, demonstrated higher correlation in the levels of data manipulation. Dumas and Devine (2000)

[3] Increasing one order of magnitude is multiplying the number by 10, increasing two orders is multiplying by 10^2, and increasing N orders of magnitude means multiplication of number by10^N. For example, 436 is one order of magnitude greater than 43.6 and is three orders of magnitude smaller than 436000.

posited that Benford’s Law, could be employed by environmental regulators to detect fraud in self-reported pollution emissions data. Benford’s Law was applied to North Carolina air pollution emissions data in an empirical example. They found that distortion in the data, which reduced the mean of the reported data by 9.5-10%, was produced because of a lack of certain first digits existed among the large firms and similarly a low frequency of another digit existed among small firms. Based on the analysis, it was concluded that smaller firms may be distorting values downward to avoid classification in the highest emission category (Title V or over 100tons/year); therefore, Benford’s Law may actually be more useful in identifying the likelihood of fraudulent reporting within categories of data instead than within specific datasets that conform or lack conformance to the Law. The study also demonstrated that if regulatory agencies adopted Benford’s Law as an auditing tool, it would still be possible to self-report fraudulent values consistent with Benford’s Law for instance by multiplying the reported values by a consistent percentage to reduce emissions of the parameters, but the use of Benford’s Law by other means would be more difficult. Subtraction of a given amount or reduction to meet a certain threshold would likely be detected by the Benford’s Law analysis.

Zahran et al. (2014) applied Benford’s Law to self-reported lead (Pb) emissions in order to evaluate the accuracy of these self-reported datasets in light of the EPA’s 2001 Final Rule governing oversight of lead emissions. The goal of their study was to identify systematic changes in firm behavior through the application of Benford’s Law to detect statistical anomalies in large datasets. The expectation was to find more inaccuracies following the new rule, which lowered the threshold for Pb emissions, but their results showed improved accuracy of self-reported Pb emissions. The study provided further evidence of the utility of statistical analysis using Benford’s Law as an auditing tool to enhance EPA regulations.

3.2 OBJECTIVES OF THE STUDY

The existing literature establishes the overall success of using Benford's Law to find irregularities in environmental datasets. Though the application of Benford's Law may not be considered a completely versatile method due to limitations on the types of data it may be applied to (i.e. large, non-uniform, unbounded, etc.), the literature has yet to establish whether it may be employed at a screening level in a tiered approach to assess the veracity of self-reported emissions data; which type of environmental data should be expected to conform to Benford's Law based on the general standards put forth for its use (Dumas & Devine, 2000). In order for Benford's Law to enhance environmental regulatory enforcement, inherent irregularities in datasets, which do not follow the law, must be distinguished from statistical anomalies caused by fraudulent or mishandled self-reported datasets.

In this study, we evaluate self-reported discharge data from wastewater treatment plants from one state agency with Benford's Law for determining:

1) The suitability of wastewater treatment plant data, which contains a variety of physical, chemical and biological parameters for assessment with Benford's Law
2) The level of conformity that suitable parameter datasets have to Benford's Law

3.3 METHODOLOGY

To achieve the objectives of the study, a dataset was obtained from a state environmental agency, containing three years of facility reported discharge parameters including data previously identified as being fraudulently reported. It consisted of 223 facilities, 354 permits, and 96 parameters. Parameters consisted of a number of permitted water quality indicators like dissolved Oxygen (DO), hardness, temperature, pH, fecal bacteria, nutrients (e.g. ammonia, sulfate, phosphorus), metals, minerals (e.g. antimony, magnesium, potassium), and herbicides (e.g. 2, 4, 6-Trichlorophenol). Choosing three years of self-reported data for the analysis could be a reasonable segment of data, as compliance monitoring frequency through inspections occurs yearly, biannually or as infrequently as every five years, based on the size of the facilities. However, the dataset has some unreported values, which may limit the analytical process. There were 4,095 total combinations of facilities and parameters. Each combination is equivalent to a single dataset for evaluation by Benford's Law.

3.3.1 Screening the Datasets for Applicability

The standards or criteria for using Benford's Law should be met in order to produce reliable results from the assessment. Therefore, some parameters and datasets should be eliminated from the analysis based on characteristics, which deem them inappropriate for Bendford's Law. The first screening involved exclusion of the parameters with a built-in minimum and maximum (like pH). The second screening consisted of assessing the uniformity of reported values for the remaining dataset. Evaluation was done based on the range between minimum and maximum of reported values for parameters of every facility and those, which were not spread across at least one order of magnitude, were eliminated from the dataset. The third screening involved elimination of very small datasets. In this study, datasets with less than

21 reported values were excluded because smaller datasets could not follow Benford's Law for the reason that follows. There should be at least 1 reported value with 9 as the first digit – this is associated with the lowest probability of appearance in Benford's Law distribution (4.6%). The minimum number of reported values with first digits between 8 to 1, were then calculated based on the rounded ratio of their percentages in Benford's Law distribution divided by 4.6. Accordingly, the minimum number of reported values for first digits 8, 7, 6, 5, 4, 3, 2 and 1 are 1, 1, 1, 2, 2, 3, 4, and 6, respectively. In the final screening step, datasets containing parameters with means less than their medians and lacking positive skewness were excluded from the evaluation set of data. After excluding the parameters that did not follow the criteria described above, the remaining self-reported datasets were tested to determine if they conform to Benford's Law.

3.3.2 Analysis

A computer code was developed in MATLAB (Statistics Toolbox Release 2014b) to fit each dataset to Benford's Law. The Pearsonian Chi-square test was calculated based on Equation 3 to evaluate the goodness of fit.

$$X^2 = N \sum_{k=1}^{9} \frac{(P(o)-P(e))^2}{P(e)} \qquad (3)$$

where, *N* is the sum of the observed frequencies, *P(o)* is the percentage of observed data and *P(e)* is expected percentage or Benford's Law distribution. Therefore, the Chi-square test provides an overall measure of the statistical deviation from the expected Benford's Law distribution of numbers in comparison with the observed distributions in the wastewater treatment discharge datasets. To establish goodness of fit, the P-value of the calculated X^2 test statistic with degrees of freedom equal to 8 must be greater than the level of significance, α. A value of α at 0.05 was used to assess the conformance of datasets with Benford's Law with 95% confidence. A more conservative confidence level of 99%, which reduces the probability of a

Type I error to 0.01, was also used in the evaluation to relax the stringency of the Chi-square goodness of fit test, which is the most used test in Benford’s Law assessments, but may be penalize certain types of datasets too harshly (Lesperance, Reed, Stephens, Tsao, & Wilton, 2016).

3.4 RESULTS

3.4.1 Parameters Eliminated Through Benford's Law Screening

The applicability of Benford's Law to wastewater treatment plant discharge data was evaluated based using the screening procedures described above. After exclusion of the datasets that did not meet the screening criteria, 690 facility/parameter combinations of 4095 remained or 17% of the initial dataset. Of the 96 reported parameters across the facilities in this dataset, only 21 parameters were able to be considered for further analysis.

To meet the criterion of excluding the parameters which have inherent minimums and maximums from the dataset, facility/parameter combinations for 6 parameters pH, pH maximum, pH minimum, Bypass Total hours per day, Dissolved Oxygen and Water Temperature were eliminated from dataset. As a result, 4095 combinations were reduced to 3473. In this step, about 90% of the excluded facility datasets parameter were associated with removing the 3 parameters - pH, Dissolved Oxygen, and Water Temperature.

The second screening process consisted of eliminating combinations in which reported values of parameters were not spread across at least one order of magnitude. As a result, 1129 of the 3473 remaining after the first screening were selected for further investigation. The majority of excluded combinations in this step consisted of flow rate, overflow occurrence and Chlorine total residual which were uniform reported values.

The third screening process was excluding combinations with a number of reported values less than 21. An additional 369 datasets were eliminated to leave 760 out of 1129. In this step, there were various parameters which were excluded from dataset. This is further described in the section 3.4.2 on classes of facilities eliminated through Benford's Law criteria screening.

The remaining combinations were screened to meet the criterion of positive skewness datasets without a mean greater than the median were excluded. The results of this step included the removal of an additional 70 combinations mostly CBOD5. Therefore, the total datasets remaining after elimination was 690, which were tested using Benford's Law to assess conformance. Table 3 contains a listing of the 75 parameters that did not meet the necessary criteria for Benford’s Law.

Table 3- Parameters Excluded After Initial Screening the Dataset – Unlikely to Fit Benford’s Law

1,4-Dichlorobenzene	Bypass Occurrence	Cyanide, Free	Mercury, Total Recoverable	Phosphorus, Total In Sludge
2,3,7,8'-TCDD TTE, Total in Sludge	Bypass Occurrence, Number per month	Cyanide, Total	Molybdenum In Sludge	Potassium In Sludge
2,4,6-Trichlorophenol	Bypass Total Hours Per Day	DDE, Whole Sample	Nickel, Total In Sludge	Salmonella Sp.
48 Hour Acute Pimephales promelas	Bypass Volume	Dieldrin, Whole Sample	Nickel, Total Recoverable	Selenium, Total In Sludge
48-Hr. Acute Toxicity Ceriodaphnia dubia	Cadmium, Total In Sludge	Dissolved Oxygen	Nitrogen Kjeldahl, Total	Selenium, Total Recoverable
7-Day Chronic Toxicity Ceriodaphnia dubia	Cadmium, Total Recoverable	Flow Rate	Nitrogen Kjeldahl, Total In Sludge	Silver, Total Recoverable
7-Day Chronic Toxicity Pimephales promelas	CBOD 5 day	Fluoranthene	Nitrogen, Inorganic, Total	Sludge Solids, Percent Total
96-Hr. Acute Toxicity Pimephales promela	Chemical Oxygen Demand (Low Level)	Gamma-BHC, Total	Oil and Grease, Freon Extr-Grav Meth	Sludge Solids, Percent Volatile
Acute Toxicity, Ceriodaphnia dubia	Chlorine, Total Residual	Heptachlor Epoxide	Oil and Grease, Hexane Extr Method	Sludge Volume, Gallons
Acute Toxicity, Pimephales promelas	Chromium, Dissolved Hexavalent	Iron, Suspended (Fe)	Oil and Grease, Total	Solids, Dissolved-Sum of
Antimony, Total	Chromium, Hexavalent (Cr +6)	Lead, Total Recoverable	Overflow Occurrence	Strontium, Total Recoverable
Antimony, Total Recoverable	Chromium, Total In Sludge	Manganese, Suspended (Mn)	Overflow Volume	Thallium, Total (TL)
Beryllium, Total In Sludge	Chromium, Total Recoverable	Mercury, Total (Hg)	pH	Thallium, Total Recoverable
Bis(2-ethylhexyl) Phthalate	Chronic Toxicity, Ceriodaphnia dubia	Mercury, Total (Low Level, PQL=1000)	pH, Maximum	Water Temperature
Bypass Duration, Hours per month	Chronic Toxicity, Pimephales promelas	Mercury, Total In Sludge	pH, Minimum	Zinc, Total In Sludge

3.4.2 Class of Facilities Eliminated Through Benford’s Law Criteria Screening

In the previous description of the screening, each criterion was addressed sequentially so that the remaining facility/parameter combinations after each screening step was the starting total for the next screening step. In this section, it is more important to know how each screening step independently impact the facilities represented in the remaining dataset, so each criterion is addressed individually based on initial number of combinations which was 4095. For example, it would be informative to know the percentage of small and large facilities eliminated due to reported parameters with inherent minimums and maximums. These results provide an understanding of what percentage of the reported values for the overall dataset by facility size consists of parameters with inherent minimums and maximums. The facility size that has the most remaining parameters compared to other categories of facilities would represent an ideal class of facilities for Benford’s Law evaluation. After the impacts of each screening criteria identified, the total impact all criteria were considered to identify the most suitable category of facilities for Benford’s Law evaluation.

Facilities were classified into 4 classifications of wastewater treatment plants by flow rate, where Class A $\geq 5\ MGD$, $1\ MGD \leq$ Class B $\leq 5\ MGD$, $100{,}000\ gpd \leq$ Class C $\leq 1\ MGD$, and Class D $\leq 100{,}000\ gpd$ (Pennsylvania Department of Environmental Protection, 2016). Higher discharge flow rate corresponds to larger facility size and the potential for more significant pollutant discharges and associated fees.

While the total facility/parameter combinations of datasets was 4095, because of missing flow rates in the reported values for three facilities, the total combinations evaluated in this study decreased to 4010.

Of the 4010 datasets, 42% of the combinations were from Class D facilities, 31% Class C, 20% Class B and 7% Class A.

To meet the first criterion, facilities with 6 reported parameters pH, pH maximum, pH minimum, Bypass Total hours per day, Dissolved Oxygen and Water Temperature were eliminated from dataset. As a result, 3407 out of 4010 facility/parameter combinations remained. Additionally, it was interesting to know that each class had almost close percentage of elimination. Results of remaining and eliminated percentage of each class are presented in Table 4.

Table 4- Results of exclusion of the parameters with inherent minimum and maximum

	Initial number of facility/paramet er combinations	**Number of combinations - After exclusion of parameters with inherent minimum and maximum (only)**	**Remaining (%)**	**Eliminated (%)**
Total in classes	4010	3407	85	15
Class A	278	247	89	11
Class B	797	709	89	11
Class C	1254	1114	89	11
Class D	1681	1337	80	20

It was also important to know which classes would be removed more from dataset if only the criterion of having reported values spread across at least one order of magnitude is met. Results of this process are presented in Table 5. Only 32% of Class D remained after this screening process. For other classes, Class C and Class B had 27% remaining; the lowest percentage of among 4 classes while Class A with 36% remaining in this step had the highest percentage. The reason could be due to more variability in the reported values of larger facilities which is normally expected.

Table 5- Results of exclusion of the parameters without at least one order of magnitude

	Initial number of facility/parameter combinations	**Number of combinations - After exclusion of parameters without at least one order of magnitude (only)**	**Remaining (%)**	**Eliminated (%)**
Total in classes	4010	1192	30	70
Class A	278	101	36	64
Class B	797	215	27	73
Class C	1254	338	27	73
Class D	1681	538	32	68

Another informative screening could help to know which sizes of facilities have the most and the least exclusion based on the number of reported values. The results could help to target certain sizes of facilities in Benford's Law evaluation with more degrees of confidence in finding self-report data mishandling. Results of exclusion of the combinations with number of reported values less than 21 are presented in Table 6. Among 4 classes, Class A had the least percentage of exclusion which was about 46%. Class B, Class C and Class D had 65%, 63% and 58% removal, respectively and elimination of 64% for Class A, 73% for Class B, 73% for Class C and 68% for Class D. As it was expected, because larger facilities report more frequently, number of reported values for these types of facilities is normally more than other facilities. Hypothesis is that Class D would have lowest number of reported values, but it does not have most elimination from datasets. The last criterion was exclusion of combinations which did not have parameters with mean greater than the median and positive skewness. Results are presented in Table 7. Class A had the least exclusion percentage, and it can be concluded that Benford's Law evaluation would have more confident in finding mishandling in wastewater treatment plants self-report data if it targets larger facilities.

Table 6- Results of exclusion of the parameters with number of reported values less than 21

	Initial number of facility/parameter combinations	Number of combinations - After exclusion of parameters with number of reported values less than 21 (only)	Remaining (%)	Eliminated (%)
Total in classes	4010	1608	40	60
Class A	278	151	54	46
Class B	797	278	35	65
Class C	1254	468	37	63
Class D	1681	711	42	58

Table 7- Results of exclusion of the parameters without mean greater than median and positive skewness

	Initial number of facility/parameter combinations	Number of combinations - After exclusion of parameters without mean greater than median and positive skewness (only)	Remaining (%)	Eliminated (%)
Total in classes	4010	2004	50	50
Class A	278	158	57	43
Class B	797	357	45	55
Class C	1254	561	45	55
Class D	1681	928	55	45

The overall number of combinations which met all criteria of Benford's Law are presented in Table 8. Class A combinations decreased from 278 to 65, which had the maximum remaining percentage among all 4 classes with 23%. Remaining percentages of Class B, Class C and Class D were 15%, 18% and 17%, respectively. Since in each screening process, Class A had less exclusion compared to other classes, this was expected to be observed in overall as well.

Table 8- Results of overall screening process

	Initial number of facility/parameter combinations	Number of combinations - After overall screening process	Remaining (%)	Eliminated (%)
Total in classes	4010	690	17	83
Class A	278	65	23	77
Class B	797	119	15	85
Class C	1254	226	18	82
Class D	1681	280	17	83

Larger facilities seem to have more suitable datasets for meeting Benford's Law criteria, so it could be helpful for environmental regulators to focus more on larger facilities when applying Benford's Law to assess the veracity of self-reported datasets.

3.4.3 Conformance to Benford's Law

After fitting each facility/parameter combination to Benford's Law, only 31% of the 690 combinations conformed the with 95% confidence. Due to the low percentage of conformity at α value equal to 0.05, the datasets were retested at an α value equivalent to 0.01 or a more conservative 99% confidence level. The results of this evaluation were an overall increase in conformance from 31% to 42%. While the percentage of the conforming datasets is low, it is highly unlikely that 58% of the self-reported datasets are fraudulent or mishandled. According to the state agency providing the raw data, only one parameter/facility combination was known to contain false data during the time period evaluated in this study. Therefore, it is more probable that characteristics of the wastewater treatment plant discharge datasets exists, which make Benford's Law unsuitable as the sole means of fraud detection in datasets of similar size and scope. Since it is well established that datasets with more reported values, that can spread across several orders of magnitudes, are more likely to follow Benford's Law, we evaluated the level of conformance based on the size of each facility/parameter data stream. Table 9 shows the level of conformance for categorized datasets at 95% and 99% confidence based on the number of reported values for every facility/parameter combination. Facility/parameter datasets were grouped into 12 categories as described in the Table 9. The count column indicates the number of facility/parameter combination datasets in each range for the number of reported values.

Although higher percentages of conformance were expected for larger datasets, this was not observed. In fact, the highest level of conformance was for the group with between 21 and 50 reported values and the lowest level of conformance was 0 for categories of reported values between 1000 to 1500, 1500 to 2000 and more than 2000. Because a consistent pattern was not observed across the categories of datasets by size, we concluded that the number of reported values or lack of reported values alone, cannot be responsible for such low conformance in this analysis. The parameters with the highest number of reported values were Total Suspended Solids, Nitrogen-Ammonia (NH_3). The parameters with the lowest number of reported values were Nitrogen-Ammonia (NH_3), Hardness-Total ($CaCO_3$), Residue-Total Dissolved, Residue-Total Filterable, Total Suspended Solids, Mercury-Total (Low Level), Nitrite Plus Nitrate-Total, Sludge-Fee Weight, Phosphorus-Total (P), Fecal Coliform, *E. coli.*, Lead-Total in Sludge, Arsenic-Total in Sludge, Barium- Total Recoverable, Copper-Total in Sludge, Copper-Total Recoverable, Zinc-Total in Sludge. Although Total Suspended Solids and Nitrogen-Ammonia (NH_3) were among both the highest and lowest numbers of reported values, the datasets with the lowest numbers of reported values and highest level of conformance to Benford’s Law also included parameters for metals and microbial indicators.

Table 9- Analysis based on number of reported values with 95% CI

Number of Reported Values (NRV)	Count	Percentage of Conforming Facilities (p=0.05)	Percentage of Conforming Facilities (p=0.01)
21<NRP<=50	323	41.18	53.25
50<NRP<=100	124	35.48	45.16
100<NRP<=150	78	17.95	26.92
150<NRP<=200	26	15.38	19.23
200<NRP<=250	23	39.36	47.83
250<NRP<=300	11	36.36	36.36
300<NRP<=350	17	17.65	17.65
350<NRP<=500	25	4.00	12
500<NRP<=1000	54	5.56	11.11
1000<NRP<=1500	5	0	0
1500<NRP<=2000	1	0	0
2000<NRP	3	0	0

As previously reported, it may be useful for a regulator to idea the likelihood of fraud among a set of self-reported values across several facilities. For this purpose, parameters were grouped into four categories - nutrients, metals, microbial indicators and solids. Table 11 shows the groupings along with the associated level of conformance.

Table 10- Analysis of Conforming Facilities based on Groupings of Parameters

Category	**Parameters**	**Percentage of conforming for each parameter (p=0.05)**	**The number of datasets out of 690 within each category**	**Percentage of Conforming Facilities (p = 0.05)**
Nutrients	Phosphorus, Total (P)	38	67	35
	Ammonia (NH3) In Sludge	17	6	
	Nitrite Plus Nitrate, Total	25	52	
	Nitrite Plus Nitrate, Total In Sludge	100	2	
	Nitrogen, Ammonia (NH3)	38	148	
Metals	Arsenic, Total In Sludge	25	5	36
	Barium, Total Recoverable	14	7	
	Copper, Total In Sludge	0	6	
	Copper, Total Recoverable	21	20	
	Lead, Total In Sludge	20	6	
	Mercury, Total (Low Level)	75	20	
	Zinc, Total Recoverable	37	19	
Microbial indicators	*E. coli*	44	33	49
	Fecal Coliform	51	55	
	Fecal Coliform in Sludge	100	2	
Solids	Hardness, Total (CaCO3)	0	24	18
	Residue, Total Dissolved	0	22	
	Residue, Total Filterable	0	14	
	Sludge Fee Weight	30	10	
	Sludge Weight	33	10	
	Total Suspended Solids	23	162	

The highest level of conformance was observed among the microbial indicators (49%), followed by the metals (36%), and nutrients (35%), which had about the same percentage of conforming datasets. Finally, the solids had the lowest level of conformance at 18%. Meanwhile, it was also noted that the range of reported values for the microbial indicators was much larger than the parameter value ranges reported for any other groups. Additionally, the range of the reported values for the parameters in the metals and nutrients were much higher than the range of reported values for solids. Hence, we found consistency with the criteria for Benford's Law in which the reported values spanning more orders of magnitude are more likely to conform to Benford's Law.

Finally, from a regulatory perspective, the identification of the types of facilities more likely to distort reporting, would be helpful for fraud detection and enforcement. Moreover, the type of facility in combination with the types of parameters, which may be fraudulently reported, would allow for more targeted investigation. The classifications of wastewater treatment plants are defined by flow rate, where $\text{Class A} \geq 5\ MGD$, $1\ MGD \leq \text{Class B} \leq 5\ MGD$, $100{,}000\ gpd \leq \text{Class C} \leq 1\ MGD$, and $\text{Class D} \leq 100{,}000\ gpd$ (Pennsylvania Department of Environmental Protection, 2016).

Table 11- Analysis of Facilities based on Classification and Grouped Parameters

Classification	Grouped parameter	Percentage of Conforming Facilities (p = 0.05)	The number of datasets out of 690 within each category
Class A	Nutrients	12.50	16
Class A	Solids	8.69	23
Class A	Microbial indicator	87.50	8
Class A	Metals	34.78	23
Class B	Nutrients	28.85	52
Class B	Solids	13.46	52
Class B	Microbial indicator	68.00	28
Class B	Metals	71.79	39
Class C	Nutrients	45.63	104
Class C	Solids	11.11	64
Class C	Microbial indicator	47.73	45
Class C	Metals	38.89	19
Class D	Nutrients	55.88	103
Class D	Solids	46.53	102
Class D	Microbial indicator	41.67	12

When the level of conformance among facilities was analyzed based on class and the grouped parameters, Table 11, no consistent pattern was observed across classes and groups. The highest percentage of facilities conforming to Benford’s Law were for Class A microbial indicators (~87%), then Class B metals (~72%) followed by Class B microbial indicators (~68%). The microbial indicator datasets had the greatest level of conformance among the four parameter groups in Class A and Class C as well. Results for Class A, Class B, Class C and Class D indicate a declining level of conformity for microbial indicators as the size of the plants are reduced-87.5%, 68%, 47.73% and 41.67%, respectively. Vidon et al. (2008) found a positive correlation between *E. coli* loads and flow rates in which higher flow rates carry higher loads of *E. coli*. A possible reason for the much higher percentage of conformity for microbial indicators in larger facilities may be the higher frequency of inspections by environmental authorities and an increased focus on these parameters. Nutrients and microbial indicators had similar levels of conformance in Class C, where only the solids (~11%) had a very low level of conformance.

Datasets for solids within Class A and B were also very low. For Class D, the smallest plants, the level of conformance was the most consistent among the groups of parameters that did not report metals.

3.4.3.1 The First Two Digits Test

The first two digits test is more robust than first digit test for revealing the violation in a dataset (Dumas & Devine, 2000). This is because of the ability of the first two digits test to gain more information in a dataset while the first digit test is hiding the violation in the mathematical basis of the law (Nigrini, 2012). First two-digit test contains the information of the second digit test as well as the information of first digit test (M. Nigrini, 2012). Comparison between first digit and first two digits tests helps to determine reliability of the test if data conforms to first two digits test, it shall also conform to first digit test. If not, Benford’s Law is not considered a reliable indicator of fraud.

Though the results of the conformity percentage by the first significant digit test were not promising, the first two-digit test was performed to compare the results in order to assess the reliability of the first digit test. As a general rule, in order to have a good fit with the first two digits of the Benford’s Law, records should not be fewer than 300 (M. Nigrini, 2012). For this purpose, the combinations of facility and parameter with less than 300 reported values, were excluded so decreased from 4095 datasets to 105 combinations. The parameters within the remaining dataset consisted of water temperature, flow rate, nitrogen-ammonia (NH_3), total suspended solids, fecal coliform, CBOD5, and chlorine-total residual.

The first step of this study, which was checking the suitability of the reported parameters was not performed before the first two digits test, since the emphasis of this part of the analysis was to present the difference between the results of first digit and first two digit tests as well as in common conforming parameters but not percentage of conformity. As it was discussed in this study, screening the parameters helps to have higher percentage of conformity and in result making more targeted decision about the non-conforming facilities, but in this case, the main goal was comparison between the performances of the tests. Furthermore, the parameters, which are not suitable for the tests will be eliminated from the conforming automatically. Among them, combination of flow rate with facilities appeared more than other combinations of parameters with facilities. The first digit test was re-run for the number of reported values greater or equal to 300. The results showed that 12 out of 105 combinations of facility/parameter conformed to the expected distribution. Based on the results above, it is clear that the number of data points are not driving the level of conformance for this type of data so this result was not completely surprising. However, only 1 out of the 105 combinations conformed to the expected distribution for the FTD test. Total suspended solid was the only parameter, which conformed to the first digit test and FTD distribution commonly. According to Nigrini (2012) probability of numbers with more digits (at least four) correlate with a good fit. Looking at the reported values for total suspended revealed that their numbers consist more digits, varying from 1 to 5 compared to other reported values, which vary from 1 to 3 digits.

3.5 DISCUSSION

Despite the limitations of Benford's Law, it is helpful in multiple cases for environmental regulators working to find fraudulent reports. Benford's Law is able to provide environmental regulators with clues to prioritize their inspection by targeting specific facility sizes and parameters. Simplicity, ease of use and the low cost of Benford's Law evaluation could always be used as an accessible tool to perform a more targeted on-site inspection in water quality monitoring. As a result, the cost of on-site inspection decreases and the risk of fraudulent self-monitoring reports could be lessened.

It should be noted that exclusion of the parameters from analysis by Benford's Law in this dataset might solely be a function of the period time represented. For example, within the 3 years dataset, there were instances where the number of reported values for certain facility/parameter combinations did not meet the minimum criteria. This is especially the case for smaller plants required to self-report quarterly unlike larger plants who report monthly. However, a period of time greater than 3 years, which would include more data points, might be considered too lengthy for the purpose of fraud detection and may be beyond the time period considered in an impact assessment

Since veracity assessment of larger facilities self-reported data with Benford's Law showed higher conformity level compared to smaller facilities, environmental regulators could inspect larger facilities more if the result of their evaluation with Benford's Law is not as expected. Furthermore, Benford's Law could be employed to assess the veracity of smaller facilities self-reported data if the provided dataset is large enough to have more variability and reported values comparable to larger facilities datasets. Providing larger datasets for smaller facilities needs to span a broader time period, unless smaller facilities are required to report more frequently. Based on these results, environmental regulators may consider requiring equal reports and inspection frequency for both large and small facilities as Weirich & Silverstein (2015) studied that permit violation is more likely in small facilities. If 50% is selected as threshold for confidence in compliance with Benford's Law for this study, the following conjectures are proposed.

Microbial indicators had the highest level of conformity with about 50% compared to other 3 groups of parameters when facility/parameters combinations were evaluated with Benford's Law. Moreover, microbial indicators had the highest level of conformity varying from 50% to 87.5% in all 4 classes of facilities, compared to other groups of parameters in Benford's Law evaluation. Therefore, microbial indicators could be considered as a more predictable and stable indicator of fraud and mishandling than other categories of parameters. As a result, environmental regulators may use microbial indicators to screen facilities or to prioritize resources for inspection. In the cases that microbial indicators, especially fecal coliforms and E. coli reported values, have a low level of conformity to Benford's Law, environmental regulators could inspect those facilities more comprehensively and frequently. Also, due to the lowest conformity level of group of solids compared to other groups of parameters, results of Benford's Law should not be considered as an indicator of fraud.

Group of Metals in Class B, and Nutrients for Class D conformed to Benford's Law with 72% and 56%, respectively. Benford's Law could be used to have more focus on Class B facilities in the cases that the conformity level of reported Metals is lower than 50%. Also, if conformity level of Nutrients for Class D is lower than 50%, there should be more frequent inspection for facilities in this category.

These results corroborate Dumas & Devine (2000) findings regarding the use of Benford's Law to determine the relative likelihood of fraudulent reports across certain types of pollutants. Also, our results are verified based on De Marchi & Hamilton (2006) discussion regarding the application of Benford's Law to self-reported data which can enable environmental regulators with a first cut on which data to trust and which data verify. This allows the EPA to determine which parameters and size of facilities should be investigated more carefully through plant inspections or monitoring.

3.5.1 Previous Applications of Benford's Law Versus Its Application in This Study

Brown (2005) applied Benford's Law to screen pollutant concentration in ambient air. The study showed that datasets with higher numbers of reported values especially with significantly high orders of magnitude had a better fit. One of the datasets was used in his study, was 24 years of reported values of 12 parameters at 17 monitoring sites. Also, other datasets were hourly measurements of one parameter in one year. In our study, since the goal was finding fraudulent reports, assessing a segment of dataset greater than three years was not reasonable due to the risks of over discharging the pollutions that may endanger aquatic life. Moreover, based on the NPDES permits, the most frequent report is daily, whereas in air quality monitoring measurement of parameters could be hourly. As a result, providing such a large dataset in wastewater treatment plants is neither feasible nor justifiable.

In the Brown (2005) study, there were some instances that the number of reported values were almost equal, but due to different types of the parameter, orders of magnitude were different. As a result, the parameters spanning a higher number of ordesr of magnitude had a better fit to Benford's Law (Durtschi et al. ,2004). This finding verifies the results of our study where datasets were evaluated based on the type of parameters, we were able to find a consistency in the levels of conformity. In result, future wastewater treatment plants self-reported datasets evaluation with Benford’s Law should focus on the type of parameters instead of number of reported values.

Brown (2005) found that the order of magnitude in datasets was larger when multiple monitoring sites or pollutant species were included in a single data set. This finding is important for the future applications of Benford’s Law in veracity assessment of wastewater treatment plants self-reported datasets, because in the screening process, elimination of parameters with order of magnitude smaller than one, resulted in decreasing 3473 facility/parameter combinations to 1129 which was about 57% of the entire dataset (4095 combinations). Including multiple parameters in each order to generate datasets spanning more orders of magnitude before Benford’s Law evaluation, could avoid a large amount of exclusion from datasets and may have more informative results. For example, all nitrogen species could be combined for each facility.

De Marchi and Hamilton (2006) assessed the accuracy of two years self-report dataset of 12 toxic release inventory chemicals. They used Benford’s Law for fraud detection like the study herein. Although the number of reported values varied from 72 to 750, which were not very different from our study, there were verified datasets from the EPA emission monitors which allowed them to confirm the accuracy of facility emissions reported values by comparison. Their first step was testing the conformance of the EPA monitoring units reported values with Benford’s Law. In the second step, testing the toxic release inventory (TRI) reported values with Benford’s Law was performed. Since the EPA reported values had a good fit with Benford’s Law distribution, they were considered as verified information to compare the veracity of toxic release inventory air emission self-reported values. Therefore, based on this study, it is not possible to know with certainty whether the self-reported values of wastewater treatment plants are fraudulent because no verified data existed, due to lack of resources and limitations of surface water monitoring. However, as it is described in the future work, current evaluation can be incorporated with verified climate and precipitation data to achieve more accurate and reliable results.

Zahran et al. (2014) applied Benford's Law to self-reported lead (Pb) emissions in order to evaluate the accuracy of these self-reported data sets in light of the EPA's 2001 Final Rule governing oversight of lead emissions. The goal of their study was to identify systematic changes in firm behavior through the application of Benford's Law to detect statistical anomalies in large datasets. They compared probability of appearance of first digit in self-reported data in 1990, 2000 and 2010 with Benford's Law. Their expectation was to find more inaccuracies following the new rule which lowered the threshold for Pb emissions, but their results showed improved accuracy of self-reported Pb emissions. Although, this study and our study both tried to discover fraudulent reports, their study goal was to track accuracy of reports with Benford's Law due to a strict change in permitted emission level of Pb at the first year of three different decades. In our study, we were trying to find conformance of facilities. The idea of tracking the changes in the accuracy of wastewater treatment plants self-reported datasets with Benford's Law in different decades may be helpful in finding fraudulent reports. This would be even more informative if the size of facilities and their permit levels have changed over time so tracking the changes in the accuracy of reports is more observable.

4 DEVELOPING A PREDICTIVE MODEL TO DETECT MISHANDLING IN THE SELF-REPORTED WATER DISCHARGE DATA

In addition to data-driven methods, multivariate statistical techniques may be useful to find irregularities in wastewater treatment plants self-reported datasets. Multivariate statistical techniques have vast applications in water resources monitoring (Shrestha & Kazama, 2007). They are robust in spatial and temporal prediction of water quality for water resources management as well as the ability of to interpret large datasets considering uncertainties.

Fannin et al. (1985) built multiple regression models using dissolved oxygen, nitrate and phosphorus as water quality parameters to estimate non-point source effects on quality of Green River of Wyoming. Fahmi et al. (2011) studied the influences of NH_3-NL and temperature on DO concentration using multiple linear regression models in Klang River, Malaysia. Tilburg & Jordan (2015) had a successful study using 12 years of data at multiple locations in the Gulf of Marine, state of Maine building multiple linear regression models for water quality prediction.

Four steps can be undertaken to develop predictive analysis

1- Correlation analysis; models that explain the variability

2- Cluster analysis in water quality based on influencing factors

3- Development of linear model

4- Validation of the models

4.1 CORRELATION ANALYSIS

Base on the available dataset, water quality parameters can be dissolved oxygen, nitrate, Biological Oxygen Demand, Total Suspended Solid, ammonia and phosphorus, which are the most frequently reported parameters in the dataset. Sampling by regulatory authorities is rare and there are no means to take parallel samples at discharge points at the same time so a verified dataset is not possible. Due to these limitations, a model must be built based on an exogenous benchmark as a metric of plausibility, which is outside the scope of the self-reported data. Then parameters were incorporated in a correlation analysis with precipitation and climate data obtained from National Climatic Data Center (NCDC) for the specific geographical locations of the facilities. Climate data would identify the seasonal variability of the parameters in the available dataset. As a sample test, correlation analysis was implemented for one facility.

According to Shapiro-Wilk test (Shapiro & Wilk, 1965), our dataset was not normally distributed so Spearman's rank correlation method, a non-parametric test was employed to find the degree of association between quantitative variables at a 0.05 alpha level. Based on the correlation analysis, causal statements like negative correlations of flow rate with water temperature and phosphorus were identified as -0.41 and -0.37 respectively. The negative correlations between flow rate and water temperature can be justified as warmer weather results in higher water temperature so it causes less precipitation, more evaporation and consequently flow rate reduction. In addition, the negative correlation of flow rate with phosphorus can be explained because of the efficiency of flow rate in nutrient removal and maintaining the water quality (Endut, Azizah, et al 2009).

4.2 CLUSTER ANALYSIS

A cluster analysis can be employed to group correlated parameters into classes based on their similarities. Singh et al. (2005) used cluster analysis in the study of Gomti river water quality assessment incorporating multivariate regression techniques. Preliminary cluster analysis for five clusters was performed. Facilities in different clusters for all correlated parameters were identified using latitude and longitude information.

4.3 DEVELOPMENT OF MULTIPLE LINEAR REGRESSION MODELS

The results of the correlation and cluster analysis can help with parameter selections for the development of linear regression models. The models could be subsequently fit using stochastic regressors through multivariate linear regression analysis that associated various water quality parameters with each other. Several multiple linear regression models could be created to predict probability distributions of dependent parameters given two or more of the independent parameters probability distributions. There have been many successful studies that used multivariate regression modeling as a tool for investigating the relationships between dependent and independent parameters in water quality prediction. Antonopoulos et al. (2001) found the existence of trends and best fitted models for nine water quality variables in eighteen years of data. Baban (1993) used regression analysis to find relationships between chlorophyll-a, total phosphorus, Secchi disk depth, suspended solids, salinity and temperature in Norfolk Broads. The result of his study showed a realistic prediction. Chen and Chang (2014) evaluated the relationship between antecedent precipitation and discharge, TSS and *E. coli* concentrations using correlation and multiple regression and found out whether and how those relationships change along an urban and rural areas.

4.4 VALIDATION OF MODELS

Validation of the models will be tested using methods of cross validation by splitting the time series across several years of self-reported data using several approaches such as 10-fold method with the entire dataset. Another approach for validation of the models can be selecting three years with the highest flow rate, and three years with the lowest flow rate, building the models and validating the models based on the rest of the dataset that has the same randomness.

Finally, splitting the times series across the years to reflect variability associated with climate change, land use and the treatment operation will be done to reflect the discrete system of surface water quality and major influencing factors. Also, establishing thresholds for consistency with predicted distributions that can be used to support decisions about the veracity of future streams of self-reported data. This is a promising approach because one would be able to define conditions like ranges of precipitation and corresponding categorized self-reported data with which to draw meaningful comparisons.

5 CONCLUSION

In this thesis, first we reviewed current monitoring processes along with identification of their gaps. Despite endeavors to monitor the water pollution discharger compliance, still there are a lot of gaps in the current programs, which are putting waterbodies in danger. Improper uncertainty analysis in TMDL program, outdated NPDES permits, fraudulent self-monitoring report, old treatment and measurement technologies in wastewater treatment plants, and gaps in inspection frequency and methods, all are the reasons for a monitoring program to be deficit. Therefore, incomplete and inaccurate generated data lessens the quality of monitoring program.

Self-monitoring reports is a method of compliance assessment, which environmental monitoring and enforcement relies extensively on regulated entities to self-report pollution discharges. Currently, the only compliance evaluation of the self-reported datasets is through visual monitoring whereas compliance evaluation of self-reported data needs a more in-depth analysis than visual evaluation. This thesis investigated a data-driven method to detect potential fraud.

As the first study, Benford's Law was used as a data-driven method because of its simplicity and ease of use. The available dataset was three years of self-reported discharge data from one state environmental agency consisting of 223 facilities, 354 permits and 96 water quality parameters.

Benford's Law was used for uncovering fraud and mishandling in wastewater treatment plants self-reported datasets in two evaluative processes. The first was evaluation of applicability of Benford's Law to the datasets by meeting the criteria. The result of this evaluation was remaining 21 parameters out of 96 parameters and 690 combinations of facility/parameters out of 4095. The next step of evaluation was analysis of the datasets for conformity to Benford's Law. Several evaluations were employed for the analysis of the dataset with first digit test; one by one of the combination of the facilities/parameters, categorizing the parameters into 12 groups based on their number of reported values, grouping the parameters into four groups of microbial indicators, nutrients, metals and solids and finally classification of facilities based on their discharge flow rate. The results showed that microbial indicators can be the group of parameters, which is more probable to conform to Benford's law. An evaluation based on first two-digit test was performed in addition to the first digit test since it is believed that mathematically FTD test can reveal some hidden information of the datasets while first digit test is unable to do. The results were surprising as only 1 parameter (TSS) out of 105 conformed to Benford's Law.

Due to the generally low conformity percentages observed across these approaches, it may be concluded that Benford's Law alone, may not be a reliable method for detecting mishandling in these types of data streams. However, the results may be useful for developing the next steps toward a data-driven method to analyze self-reported data using predictive models. It was found that focusing on microbial indicators may be a more predictable; and therefore, stable indicator of fraud and mishandling than the other categories of parameters.

With performing the FTD test, it was revealed that low conformity of the first digit test is lower than what it was thought. With the first digit test, 12 out of 105 (about 11%) conformed to the Benford's Law whereas with the FTD analysis 1 out 105 (about 1%) was in conformity with the Law. The difference in conformity to the Law with the two tests was about 10%, which is not negligible. It was perceived that the results of the first digit analysis is not reliable by itself and FTD test can be added to see how more information can be captured through it.

6 FUTURE WORKS

There were limitations to our Benford's Law analysis that needs to be addressed in future work. Nigrini (2012) suggests, as a general rule, there should be at least 1000 records to expect a good fit. The dataset evaluated over 3 years was not that large so it may be one the reasons for the low conformity. The probability of numbers with more digits (at least four) correlates with a good fit (Nigrini 2012), though the majority of the numbers in our dataset had less than two digits.

Despite our lack of success in using Benford's Law for uncovering fraud and mishandling in wastewater treatment plants self-reported datasets, the use of multivariate statistical techniques was discussed as one of the promising approaches for future studies. The models are robust in spatial and temporal prediction of water resources quality as well as the ability of to interpret large datasets considering uncertainties. Future study could consist of three steps of correlation and cluster analysis, development of linear models and validation of the models. Limitation of this approach was due to sampling by regulatory authorities, which not only is not as frequent as plants sampling, but also taking parallel samples at the same time points of the plants sampling is not possible. Therefore, this causes lack of possession of verified data and in result, building models must be based on an exogenous benchmark as a metric of plausibility which is outside the scope of the self-reported data.

APPENDIX

APPENDIX

Benford’s Law first and first two digits MATLAB codes

```
% Benford’s Law first digit test
%Meeting BL critera

clear;clc;close all;

%% loading the data

load DMR

%% categorizing the data

resultsTBL = grpstats(DMR, ...
                      {'facility','parameter'}, ...
                      {@(c) nanmean(c), @(c) sum(~isnan(c),1),
'min','max','median','skewness'}, ...
                      'DataVars','reportedValue');
resultsTBL.Properties.VariableNames(end-5:end) =
{'mean','numel','min','max','median','skewness'};
resultsTBL.Properties.RowNames = {};

%% Meeting the criteria

excludeList = {'Bypass Total Hours Per Day', ...
               'Dissolved Oxygen', ...
               'pH', ...
               'pH, Maximum', ...
               'pH, Minimum', ...
               'Water Temperature'};

resultsTBL.cond2 = ~ismember(resultsTBL.parameter,excludeList);

resultsTBL.cond3 = (resultsTBL.max ./ resultsTBL.min) >=1;
resultsTBL.cond4 = resultsTBL.numel>21;
resultsTBL.cond5 = and ( (resultsTBL.mean > resultsTBL.median),
...
                         resultsTBL.skewness>0 );

resultsTBL.allConds = all(table2array(resultsTBL(:,
{'cond2','cond3','cond4','cond5'})),2);
```

```
fprintf('Remaining (Facility,Parameter) combinations: %d \n',
sum(resultsTBL.allConds));


%% Meeting the criteria based on facility size
classNames = {'Class D', 'Class C', 'Class B', 'Class A'};
tmpDMR = resultsTBL(strcmpi(resultsTBL.parameter,'Flow
Rate'),[1,2,4]);
tmpMask = isnan(tmpDMR.mean);
tmpDMR.mean(tmpMask) = 0;
tmpDMR.classID = imquantize(tmpDMR.mean, [100e3, 1e6, 5e6]);
tmpDMR.class = cell(size(tmpDMR,1),1);
tmpDMR.class(:) = classNames(tmpDMR.classID);
tmpDMR.classID(tmpMask) = NaN;
tmpDMR.class(tmpMask) = {''};

resultsTBL = outerjoin(resultsTBL, tmpDMR, ...
                  'keys', {'facility'}, ...
                  'RightVariables',{'class'});
clear tmpDMR tmpMask
writetable(resultsTBL,'resultsTBL.xlsx')
resultsTBL2 = grpstats(resultsTBL, ...
                       {'class','parameter'}, ...
                       {'sum'}, ...

'DataVars',{'cond2','cond3','cond4','cond5','allConds'});
resultsTBL2.Properties.RowNames = {};
writetable(resultsTBL2,'resultsTBL2.xlsx');
resultsTBL3 = grpstats(resultsTBL, ...
                       {'class'}, ...
                       {'sum'}, ...

'DataVars',{'cond2','cond3','cond4','cond5','allConds'});
resultsTBL3.Properties.RowNames = {};
writetable(resultsTBL3,'resultsTBL3.xlsx');



% Calculation of the percentage and plotting the results versus
BL

clc
clear all

%% Loading Data

dataTBL=readtable('EPADischargeData.csv','HeaderLines',1);
```

```
%% Finding non-NaN mask for reported values

mask=isnan(dataTBL.ReportedValue);

%% Extracting the most significant digit for reported Value

dataTBL.MSD=NaN(size(dataTBL,1),1);
dataTBL.MSD(~mask)=arrayfun(@(v)
MSD1(v),dataTBL.ReportedValue(~mask));
ResultsTBL=grpstats( dataTBL,{'Facility','Parameter'}, ...
                     {@(v)
sum(v(~isnan(v))==1)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==2)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==3)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==4)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==5)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==6)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==7)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==8)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==9)./numel(v(~isnan(v))), ...
                      @(v) numel(v(~isnan(v)))},...
                     'DataVars',{'MSD'});




%% Plotting

for i=1:size(ResultsTBL,1)
    NumberOfData=ResultsTBL(i,13);
    figure
    bar([table2array(ResultsTBL(i,4:12))',E]*100);
    ylim([0 100]);
    legend('EPA Discharge Data','Benford''s Law Distribution');
    title(sprintf('%s -
%s',cell2mat(ResultsTBL.Facility(i)),cell2mat(ResultsTBL.Paramet
er(i))))
    xlabel('MSD')
    ylabel('Percentage')
    saveas(gcf,sprintf('fig%d.png',i),'png');
    close

end

%% Statistical Analysis

results2=table();
results2.N=ResultsTBL.Fun10_MSD;

results2.Chi2=ResultsTBL.Fun10_MSD.*sum((table2array(ResultsTBL(
:,4:12))-
repmat(benfords,nRows,1)).^2./(repmat(benfords,nRows,1)),2;

%% Chi-square formula

results2.p=1-chi2cdf(results2.Chi2,8);

results2.mask=double(results2.p>0.05);
results2.mask(isnan(results2.p))=NaN;


for i=1:size(results2,1)
  if results2.mask(i)==0
    results2.conformity(i)={'Does not conform'};
  elseif isnan(results2.mask(i))
    results2.conformity(i)={'Not a Number'};
  elseif results2.mask(i)==1
    results2.conformity(i)={'Conforms'};
  else
    error('Mask Not recognized')
  end
end



%% Grouping based on number of reported values

edges=[50,100,150,200,250,300,350,500,1000,1500,2000];
groupID=imquantize(results2.N,edges);
countPerGroup=arrayfun(@(id) sum(groupID==id),
(1:numel(edges)+1)');
percentConformalPerGroup=arrayfun(@(id)
sum(results2.mask(groupID==id))./countPerGroup(id)*100,
(1:numel(edges)+1)');
fprintf('%5s  %6s  %10s\n','Group','Count','%')
fprintf('%5g  %6g
%10.2f\n',[(1:numel(edges)+1);countPerGroup';percentConformalPer
Group'])

edges=[50,100,150,200,250,300,350,500,1000,1500,2000];
nGroup=numel(edges)+1;
groupID=imquantize(results2.N,edges);
percentConformalPerGroup=zeros(nGroup,1);
for id=1:nGroup
  tmpMask=groupID==id;
  countPerGroup(id)=sum(tmpMask);

percentConformalPerGroup(id)=sum(results2.mask(tmpMask))./countP
erGroup(id)*100;
end
fprintf('%5s  %6s  %10s\n','Group','Count','%')
fprintf('-------------------------\n')
fprintf('%5g  %6g
%10.2f\n',[(1:numel(edges)+1);countPerGroup';percentConformalPer
Group'])
fprintf('-------------------------\n')

%% Calculation of conformity percentage

conformity=sum(results2.mask(:) == 1);
ConformityPercentage=conformity./size(results2,1);

%% Percentage of conformity for each parameter

PP =
table(ResultsTBL.Facility,ResultsTBL.Parameter,results2.mask);

individual_Parameter_Conformity=grpstats(PP,{'Var1','Var2','Var3
'});




%% Plotting

for i=1:size(ResultsTBL,1)
  if (~strcmpi(results2.conformity{i},'Not a Number'))
    figure
    bar([table2array(ResultsTBL(i,4:12))' benfords']*100);
    ylim([0 100]);
    legend('EPA Discharge Data','Benford''s Law Distribution');
    title(sprintf('%s - %s -
%s',ResultsTBL.Facility{i},ResultsTBL.Parameter{i},results2.conf
ormity{i})) %
    xlabel('MSD')
    ylabel('Percentage')
    saveas(gcf,sprintf('fig%d.png',i),'png');
    close
  end
end

% clc
% clear all
%% Loading Data

% Grouping based on the group of parameters and facility size

dataTable=readtable('wholedata.csv','HeaderLines',1);

for i=1:size(dataTable,1)
    i
    A(i)=strcmp(dataTable.WaterTemperature(i),'Flow Rate');
end
B=find(A==0);
dataTable(B,:)=[];

save('Data2','dataTable')

% % load('Data2')
for i=1:size(dataTable,1)
    i
    if strcmp(dataTable.C(i),'CFS')
        dataTable.Var7(i)=dataTable.Var7(i)*0.646316889697*10^6;
%           dataTable.C(i)='GPD';
    elseif strcmp(dataTable.C(i),'MGD')
        dataTable.Var7(i)=dataTable.Var7(i)*10^6;
%           dataTable.C(i)='GPD';
    end
end

ER=isnan(dataTable.Var7);
ER2=find(ER==1);
dataTable(ER2,:)=[];
dataTable2=dataTable(:,{'JellowayWWTP','Var7'});
F=grpstats(dataTable2,'JellowayWWTP');

G=F.mean_Var7;
GD=F.JellowayWWTP;
```

```
ClassD=find(G < 100000);
NameD=GD(ClassD);
G(ClassD)=[];
GD(ClassD)=[];
ClassC=find( G < 1000000);
NameC=GD(ClassC);
G(ClassC)=[];
GD(ClassC)=[];
ClassB=find( G < 5000000);
NameB=GD(ClassB);
G(ClassB)=[];
GD(ClassB)=[];
ClassA=find( G > 5000000);
NameA=GD(ClassA);
G(ClassA)=[];
GD(ClassA)=[];

for y=1:size(F,1)
    y
    if F.mean_Var7(y)<=100000
        F.class(y)={'class D'}
    elseif F.mean_Var7(y)<=1000000
        F.class(y)={'class C'}
    elseif F.mean_Var7(y)<=5000000
        F.class(y)={'class B'}
    else
        F.class(y)={'class A'}
    end
end

%  Nutrients={'Ammonia (NH3) In Sludge',...
         'Nitrite Plus Nitrate, Total',...
         'Nitrite Plus Nitrate, Total In Sludge',...
         'Phosphorus, Total (P)'...
         'Nitrogen, Ammonia (NH3)'};


validParamMaskNutr=ismember(ResultsTBL.Parameter,Nutrients)
for u=1:size(ResultsTBL,1)
    if validParamMaskNutr(u)==1
        ResultsTBL.ParameterGRP(u)={'Nutrients'}
    end
end
%%
Metals={'Arsenic, Total In Sludge',...
        'Barium, Total Recoverable',...
        'Copper, Total In Sludge',...
        'Copper, Total Recoverable',...
        'Lead, Total In Sludge',...
         'Mercury, Total (Low Level)',...
         'Zinc, Total Recoverable'};
validParamMaskMetl=ismember(ResultsTBL.Parameter,Metals)
for r=1:size(ResultsTBL,1)
    if validParamMaskMetl(r)==1
        ResultsTBL.ParameterGRP(r)={'Metals'}
    end
end
%%
Microbial_indicator={'E. coli', ...
           'Fecal Coliform', ...
           'Fecal Coliform in Sludge'};
validParamMaskMicrobindc=ismember(ResultsTBL.Parameter,Microbial
_indicator)
for e=1:size(ResultsTBL,1)
    if validParamMaskMicrobindc(e)==1
        ResultsTBL.ParameterGRP(e)={'Mictobial indicator'}
    end
end
%%
Solids={'Hardness, Total (CaCO3)', ...
            'Residue, Total Dissolved', ...
            'Residue, Total Filterable', ...
            'Sludge Fee Weight', ...
            'Sludge Weight', ...
             'Total Suspended Solids'};
validParamMaskSolid=ismember(ResultsTBL.Parameter,Solids)

for w=1:size(ResultsTBL,1)
    if validParamMaskSolid(w)==1
        ResultsTBL.ParameterGRP(w)={'Solids'}
    end
end

 clc
clear all

%% First Two digits of Benford's Law analysis
%% Loading Data
dataTBL=readtable('EPADischargeData.csv','HeaderLines',1);

%% Finding non-NaN mask for reported values
mask=isnan(dataTBL.ReportedValue);

%% getting first two digits for reported Value
```

```
dataTBL.FTD=NaN(size(dataTBL,1),1);
dataTBL.FTD(~mask)=arrayfun(@(v)
MSDft(v),dataTBL.ReportedValue(~mask));

%% Categorizing the data based on facility and parameters and
calculating the percentage of FTD of each reported value
ResultsTBL=grpstats( dataTBL,{'Facility','Parameter'}, ...
                     {@(v)
sum(v(~isnan(v))==10)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==11)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==12)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==13)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==14)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==15)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==16)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==17)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==18)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==19)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==20)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==21)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==22)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==23)./numel(v(~isnan(v))), ...
                      @(v)
sum(v(~isnan(v))==24)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==25)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==26)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==27)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==28)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==29)./numel(v(~isnan(v))), ...
```

```
                     @(v)
sum(v(~isnan(v))==30)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==31)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==32)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==33)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==34)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==35)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==36)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==37)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==38)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==39)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==40)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==41)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==42)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==43)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==44)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==45)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==46)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==47)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==48)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==49)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==50)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==51)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==52)./numel(v(~isnan(v))), ...
```

```
                     @(v)
sum(v(~isnan(v))==53)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==54)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==55)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==56)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==57)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==58)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==59)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==60)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==61)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==62)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==63)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==64)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==65)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==66)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==67)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==68)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==69)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==70)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==71)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==72)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==73)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==74)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==75)./numel(v(~isnan(v))), ...
```

```
                     @(v)
sum(v(~isnan(v))==76)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==77)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==78)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==79)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==80)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==81)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==82)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==83)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==84)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==85)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==86)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==87)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==88)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==89)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==90)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==91)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==92)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==93)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==94)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==95)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==96)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==97)./numel(v(~isnan(v))), ...
                     @(v)
sum(v(~isnan(v))==98)./numel(v(~isnan(v))), ...
```

```
                     @(v)
sum(v(~isnan(v))==99)./numel(v(~isnan(v))), ...
                      @(v) numel(v(~isnan(v)))},...
                     'DataVars',{'FTD'});



validParamMask=ismember(ResultsTBL.Parameter,validParameters);
ResultsTBL=ResultsTBL(validParamMask,:);
%% Calculating the percentage of BL formula for the first two
digits
p=deal([]);
for i=10:99
    digits=dec2base(i,10) - '0'
        prob=log10(1+1/(10*digits(1,1)+digits(1,2)));
     p=[p,prob];

end
p=p';
a=(10:99);
a=a';
b=[a,p];



toDelete = ResultsTBL.Fun91_FTD<300;
ResultsTBL(toDelete,:) = [];
nRows=size(ResultsTBL,1);

%% Statistical Analysis
results2=table();
results2.N=ResultsTBL.Fun91_FTD;

results2.Chi2=ResultsTBL.Fun91_FTD.*sum((table2array(ResultsTBL(
:,4:93))-repmat(p,nRows,1)).^2./(repmat(p,nRows,1)),2);

results2.p=1-chi2cdf(results2.Chi2,89);
results2.mask=double(results2.p>0.05);


for i=1:size(results2,1)
  if results2.mask(i)==0
    results2.conformity(i)={'Does not conform'};
  elseif isnan(results2.mask(i))
    results2.conformity(i)={'Not a Number'};
  elseif results2.mask(i)==1
    results2.conformity(i)={'Conforms'};
  else
    error('Mask Not recognized')
  end
end



%% Plotting
for i=1:size(ResultsTBL,1)
plot(a,p*100,'r','linewidth',1);
grid on
grid minor
axis([10 100 0 15])
title('FIRST TWO DIGITS ANALYSIS OF BENFORDS LAW')
xlabel('First two digits')
ylabel('Proportion')
hold on
x=10:99;
y=[table2array(ResultsTBL(i,4:93))]*100;
bar(x,y);
legend('Benford''s Law First Two digits Distribution','EPA
Discharge Data');
title(sprintf('%s -
%s',cell2mat(ResultsTBL.Facility(i)),cell2mat(ResultsTBL.Paramet
er(i))))
saveas(gcf,sprintf('fig%d.png',i),'png');
close

end
```

REFERENCES

REFERENCES

Alparslan, E., Aydöner, C., Tufekci, V., & Tüfekci, H. (2007). Water quality assessment at Ömerli Dam using remote sensing techniques. *Monitoring and Assessment*.

Andreen, W. L. (2004). Water Quality Today - Has the Clean Water Act Been a Success? *Alamaba Law Reivew*, *55*, 537–593.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572.

Birkeland, S. (2001). EPA’s TMDL Program. *Ecology Law Quarterly*, *28*, 297–305.

Boyd, C. E. (2003). Guidelines for aquaculture effluent management at the farm-level. In *Aquaculture* (Vol. 226, pp. 101–112).

Brown, R. J. C. (2005). Benford’s Law and the screening of analytical data: the case of pollutant concentrations in ambient air. *Analyst*, *130*(9), 1280–1285.

Caccia, V. V. G., & Boyer, J. N. J. (2005). Spatial patterning of water quality in Biscayne Bay, Florida as a function of land use and water management. *Marine Pollution Bulletin*, *50*(11), 1416–1429.

Carslaw, C. A. P. N. (1988). Anomalies in income numbers: Evidence of goal oriented behavior. *Accounting Review*, 321–327.

Copeland, C. (2010). Clean Water Act : A Summary of the Law. *Water*, *2011*(October 24).

Copeland, C. (2014). Clean water act: Asummary of the law. In *Oil and Chemical Spills: Federal Emergency Response Framework and Related Legal Authorities* (pp. 29–44).

Davidson, E. A., David, M. B., Galloway, J. N., Goodale, C. L., Haeuber, R., Harrison, J. A., … Ward, M. H. (2011). Excess nitrogen in the U.S. environment: Trends, risks, and solutions. *Issues in Ecology*, (15).

De Marchi, S., & Hamilton, J. T. (2006). Assessing the accuracy of self-reported data: an evaluation of the toxics release inventory. *Journal of Risk and Uncertainty*, *32*(1), 57–76.

Deily, M., & Gray, W. (1991). Enforcement of pollution regulations in a declining industry. *Journal of Environmental Economics.*

Dilks, D., & Freedman, P. (2004). Improved consideration of the margin of safety in total maximum daily load development. *Journal of Environmental Engineering*.

Docampo, S., del Mar Trigo, M., Aira, M. J., Cabezudo, B., & Flores-Moya, A. (2009). Benford’s law applied to aerobiological data and its potential as a quality control tool.

Aerobiologia, *25*(4), 275–283.

Doremus, H., & Dan Tarlock, A. (2012). Can the U.S. Clean water act succeed as an ecosystem protection law? *Journal of Water Law, 23*(1), 3–23.

Drake, P. D., & Nigrini, M. J. (2000). Computer assisted analytical procedures using Benford's law. *Journal of Accounting Education, 18*(2), 127–146.

Duhigg, C. (2009). Sewers at Capacity, Pollution Spills Into Waterways -Series -NYTimes As Sewers Fill, Waste Poisons Waterways.

Dumas, C. F., & Devine, J. H. (2000). Detecting Evidence of Non Compliance in Self-Reported Pollution Emissions Data: An Application of Benford's Law. In *American Agricultural Economics Association Annual Meeting, Tampa*.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The Effective Use of Benford ' s Law to Assist in Detecting Fraud in Accounting Data. *Journal of Forensic Accounting, 99*(99), 17–34.

Earnhart, D. (2004). Panel data analysis of regulatory factors shaping environmental performance. *Review of Economics and Statistics*, *86*(1), 391–401.

Earnhart, D. (2010). Effluent Limits and Monitoring: Do Regulators Inspect Polluters Facing Tighter Limits Less Frequently in Response to Noncompliance?

EPA. (2002). *Introduction to the Clean Water Act*.

EPA. (2004). NPDES Compliance Inspection Manual, (July).

EPA. (2006). NEPIS Document display. Retrieved from http://nepis.epa.gov

EPA. (2010a). *NPDES Permit Writers' Manual*.

EPA. (2010b). U.S. Environmental Protection Agency NPDES Permit Writers' Manual.

Fahmi, M., Nasir, M., Samsudin, M. S., Mohamad, I., Roshide, M., Awaluddin, A., … Ramli, N. (2011). River Water Quality Modeling Using Combined Principle Component Analysis (PCA) and Multiple Linear Regressions (MLR): A Case Study at Klang River , Malaysia Department of Environmental Sciences , Faculty of Environmental Studies , Department of Enviro. *World Applied Sciences Journal*, *14*(2002), 73–82.

Fannin, T. E., Parker, M., & Maret, T. J. (1985). Multiple Regression Analysis for Evaluating Non-Point Source Contributions to Water Quality in the Green River, Wyoming1. *Wetlands*.

Flajsig, G. (1999). Common problems in wastewater treatment plants function. *WIT Transactions on Ecology and the Environment*.

Freeman, A. M. (1990). *Public Policies for Environmental Protection*.

Fu, Q., Fang, Z., Villas-Boas, S. B., & Judge, G. (2014). An Investigation of the Quality of Air Data in Beijing.

Gaba, J. (2007). Generally illegal: NPDES general permits under the Clean Water Act. *Harvard Environmental Law Review*.

GAO. (1983). *Wastewater Discharges Are Not Complying With EPA Pollution Control Permits*

GAO. (2007). GAO-07-731G Government Auditing Standards: July 2007 Revision. *Auditing*, (July).

Giardino, C., Brando, V., & Dekker, A. (2007). Assessment of water quality in Lake Garda (Italy) using Hyperion. *Remote Sensing of*.

Glicksman, R., & Earnhart, D. (2007). The Comparative Effectiveness of Government Interventions on Environmental Performance in the Chemical Industry. *Stanford Environmental Law Journal*, *719*(forthcoming), 1–43.

Gray, W. B., & Shimshack, J. P. (2011). The effectiveness of environmental monitoring and enforcement: A review of the empirical evidence. *Review of Environmental Economics and Policy*.

Harmancioglu, N. B., Fistikoglu, O., Ozkul, S. D., Singh, V. P., & Alpaslan, M. N. (1999). *Water Quality Monitoring Network Design*. *Environmental data*.

He, W., Chen, S., Liu, X., & Chen, J. (2008). Water quality monitoring in a slightly-polluted inland water body through remote sensing—case study of the Guanting Reservoir in Beijing, China. *Of Environmental Science & Engineering.*

Horowitz, A. (2013). A review of selected inorganic surface water quality-monitoring practices: are we really measuring what we think, and if so, are we doing it right? *Environmental Science & Technology*.

Houck, O. A. (2002). *The Clean Water Act TMDL program: law, policy, and implementation*. Environmental Law Institute.

Jury, W. A., & Vaux, H. (2005). The role of science in solving the world's emerging water problems. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(44), 15715–15720.

Kaplow, L., & Shavell, S. (1991). Optimal law enforcement with self-reporting of behavior.

Khoshelham, K. (2011). Accuracy analysis of kinect depth data. *ISPRS Workshop Laser Scanning*.

Kleit, A., Pierce, M., & Hill, R. (1998). Environmental protection, agency motivations, and rent extraction: The regulation of water pollution in Louisiana. *Journal of Regulatory Economics*.

Kossovsky, A. E. (2015). Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications.

Lesperance, M., Reed, W. J., Stephens, M. A., Tsao, C., & Wilton, B. (2016). Assessing Conformance with Benford's Law: Goodness-Of-Fit Tests and Simultaneous Confidence Intervals. *PLOS ONE*, *11*(3), e0151235.

Lillesand, T., Kiefer, R., & Chipman, J. (2014). *Remote sensing and image interpretation*.

Lovett, G., Burns, D., & Driscoll, C. (2007). Who needs environmental monitoring? *Frontiers in Ecology*.

Maillard, P., & Santos, N. (2008). A spatial-statistical approach for modeling the effect of non-point source pollution on different water quality parameters in the Velhas river watershed–Brazil. *Journal of Environmental Management*.

Miller-mcclellan, J., Shanholtz, V., & Miller-mcclellan, J. (2003). A Comparative Study of the Total Maximum Daily Load (TMDL) Program and Process in Virginia and Kansas : Possible Outcomes and Effects upon Stakeholders A Comparative Study of the Total Maximum Daily Load Possible Outcomes and Effects upon Stakeholders.

Ming-kui, Z., Li-ping, W., & Zhen-li, H. E. (2007). Spatial and temporal variation of nitrogen exported by runo ff from sandy agricultural soils. *Journal of Environmental Sciences*, *19*(3), 1086–1092.

Muñoz-Carpena, R., Vellidis, G., Shirmohammadi, A., & Wallender, W. W. (2006). Evaluation of Modeling Tools for TMDL Development and Implementation. *Transactions of the ASABE*, *49*(4), 961–965.

Nigrini, M. (1994). Using digital frequencies to detect fraud. *The White Paper*, *8*(2), 3–6.

Nigrini, M. (2007). Digital Analysis Using Benford's Law: Tests and Statistics for Auditors. *Social Sciences*, *34*(3).

Nigrini, M. (2011). *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations* (Vol. 558). John Wiley & Sons.

Nigrini, M. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection* (Vol. 586). John Wiley & Sons.

Nigrini, M. J. (1992). The detection of income tax evasion through an analysis of digital frequencies. *Doctorat En Sciences de Gestion, Cincinnati: Université de Cincinnati*.

Nigrini, M. J. (1996). A taxpayer compliance application of Benford's law. *The Journal of the American Taxation Association*, *18*(1), 72.

Nigrini, M. J. (1999). I've got your number. *Journal of Accountancy*, *187*(5), 79.

Nigrini, M. J. (2003). *Using Microsoft Access for Data Analysis and Interrogation: The Use of Benford's Law, Number Patterns, Ratios, and Duplications to Detect Errors, Biases, Fraud, Irregularities, and Inefficiencies in Corporate Data*. Mark J. Nigrini.

Nigrini, M. J. (2005). Inspiration from Beethoven's Sixth: the flawless performance of a symphonic masterpiece, and the lessons it offers, can be music to an auditor's ears. *Internal Auditor*, *62*(4), 52–57.

Nigrini, M. J., & Miller, S. J. (2007). Benford's law applied to hydrology data—results and relevance to other geophysical data. *Mathematical Geology*, *39*(5), 469–490.

Nigrini, M. J., & Miller, S. J. (2009). Data diagnostics using second-order tests of Benford's law. *Auditing: A Journal of Practice & Theory*, *28*(2), 305–324.

Nigrini, M. J., & Mittermaier, L. J. (1997). The use of Benford's law as an aid in analytical procedures. *Auditing*, *16*(2), 52.

NYSDEC. (2012). *SPDES Compliance and Enforcement - SFY2012/2013*.

Parry, R. (1998). Agricultural Phosphorus and Water Quality: A U.S. Environmental Protection Agency Perspective. *Journal of Environment Quality*, *27*(2), 258.

Peltzman, S. (1976). Toward a more general theory of regulation.

Pennsylvania Department of Environmental Protection. (2016). Definitions of Classes and Subclasses.

PUCKETT, L. J. (1995). Identifying the Major Sources of Nutrient Water Pollution. *Environmental Science & Technology*, *29*(9), 408A–414A.

Rechtschaffen, C., & Markell, D. (2003). *Reinventing environmental enforcement and the state/federal relationship*.

Ritchie, J. C., & Cooper, C. M. (1988). Comparison of measured suspended sediment concentrations with suspended sediment concentrations estimated from Landsat MSS data. *Int.J.Remote Sensing*, *9*(3), 379–387.

Rivers, L., Dempsey, T., Mitchell, J., & Gibbs, C. (2015). Environmental Regulation and Enforcement: Structures, Processes and the Use of Data for Fraud Detection. *Journal of Environmental Assessment Policy and Management*, 1550033.

Roukema, B. F. (2009). Benford's Law anomalies in the 2009 Iranian presidential election. *Unpublished Manuscript*.

Sambridge, M., Tkalčić, H., & Jackson, A. (2010). Benford's law in the natural sciences. *Geophysical Research Letters*, *37*(22).

Sanders, T. G. (1983). *Design of Networks for Monitoring Water Quality*.

Schaepman-Strub, G., Schaepman, M., Painter, T., Dangel, S., & Martonchik, J. (2006). Reflectance quantities in optical remote sensing - definitions. *Remote Sensing of Environment*, *103*, 27–42.

Schalles, J. F., Gitelson, A. a, Yacobi, Y. Z., & Kroenke, A. E. (1998). ESTIMATION OF CHLOROPHYLL a FROM TIME SERIES MEASUREMENTS OF HIGH SPECTRAL RESOLUTION REFLECTANCE IN AN EUTROPHIC LAKE 1 sured indicator of algal density and is a key param- systems . The unique light absorption pattern of Numerous studies during the last. *Water*, *390*(January 1997), 383–390.

Seyhan, E., & Dekker, A. (1986). Application of Remote Sensing Techniques for Water Quality Monitoring. *Hydrobiological Bulletin*, *20*, 41–50.

Shapiro, S., & Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*.

Shimshack, J. P., & Ward, M. B. (2005). Regulator reputation, enforcement, and environmental compliance. *Journal of Environmental Economics and Management*, *50*(3), 519–540.

Shirmohammadi, A., Chaubey, I., & Harmel, R. (2006). Uncertainty in TMDL models.

Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling and Software*, *22*(4), 464–475.

Situ, Y., & Emmons, D. (2000). *Environmental crime : the criminal justice system's role in protecting the environment*. Sage Publications.

Storey, M., Gaag, B. van der, & Burns, B. (2011). Advances in on-line drinking water quality monitoring and early warning systems. *Water Research*.

Tilburg, C., & Jordan, L. (2015). The effects of precipitation, river discharge, land use and coastal circulation on water quality in coastal Maine. *Royal Society*.

Tjeerdema, R. (2007). *Navigating the TMDL process: sediment toxicity*.

Tsakiris, G., & Alexakis, D. (2012). Water quality models: An overview. *European Water*, *37*, 33–46.

US EPA Office of Inspector General (OIG). (1999). Laboratory Fraud: Deterrence and Detection.

US EPA OIG. (2014). EPA Has Not Implemented Adequate Management Procedures to Address Potential Fraudulent Environmental Data Recommendations and Planned Corrective Actions.

US GAO. (1993). *Environmental Enforcement: EPA Cannot Ensure the Accuracy of Self-Reported Compliance Monitoring Data*.

USGS. (2008). *Report as of FY2008 for 2008LA58B: &quot;Uncertainty-based TMDL Calculations for Dissolved Oxygen in Amite River&quot; Publications Report as of FY2008 for 2008LA58B: &quot;Uncertainty-based TMDL Calculations for Dissolved Oxygen in Amite 1*.

Vidon, P., Tedesco, L. P., Wilson, J., Campbell, M. A., Casey, L. R., & Gray, M. (2008). Direct and indirect hydrological controls on concentration and loading in midwestern streams. *Journal of Environmental Quality*, *37*(5), 1761–1768.

Vries, P. de, & Murk, A. (2013). Compliance of LC50 and NOEC data with Benford's Law: An indication of reliability? *Ecotoxicology and Environmental Safety*.

Wallace, W. A. (2002). Assessing the quality of data used for benchmarking and decision-making. *The Journal of Government Financial Management*, *51*(3), 16.

Wang, X. (2001). Integrating water-quality management and land-use planning in a watershed context. *Journal of Environmental Management*, *61*(1), 25–36.

Wesley A. Magat and W. Kip Viscusi. (1990). Effectiveness of the EPA's Regulatory Enforcement: The Case of Industrial Effluent Standards. *Journal of Law and Economics*, *33*(October), 331–360.

Zahran, S., Iverson, T., Weiler, S., & Underwood, A. (2014). Evidence that the accuracy of self-reported lead emissions data improved: A puzzle and discussion. *Journal of Risk.*

Zhang, H. X., & Yu, S. L. (2004). Applying the first-order error analysis in determining the margin of safety for total maximum daily load computations. *Journal of Environmental Engineering*, *130*(6), 664–673.