The Impact of Water System Ownership on

Regulatory Compliance: A Multilevel Analysis of

Drinking Water Quality Violations in California

A Thesis Submitted to the Graduate Faculty of the National University

Department of Engineering, Data and Computer Sciences

in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

Prepared By:

William Owens

Colin Smith

Brad Morse

National University

June 2024

Master's Thesis Approval Form

We certify that we have read the project of William Owens, Colin Smith, and Brad Morse entitled THE IMPACT OF WATER SYSTEM OWNERSHIP ON REGULATORY COMPLIANCE: A MULTILEVEL ANALYSIS OF DRINKING WATER QUALITY VIOLATIONS IN CALIFORNIA and that, in our opinion, it is satisfactory in scope and quality as the thesis for the degree of Master of Science in Data Science at National University.

Approved:

Aeron Zentner, DBA, Capstone Instructor   Date
Associate Part-Time Professor
Department of Engineering, Data and Computer Sciences
School of Technology and Engineering
College of Business, Engineering and Technology
National University

Shatha Jawad, Ph.D., Capstone Project Advisor   Date
Professor
Department of Computer Science & Cybersecurity
School of Technology and Engineering
College of Business, Engineering and Technology
National University

Mohammad Yavarimanesh, Ph.D., Capstone Project Sponsor   Date
Assistant Part-Time Professor
Department of Engineering, Data and Computer Sciences
School of Technology and Engineering
College of Business, Engineering and Technology
National University

Abstract

This study investigates the relationship between water system ownership and regulatory violations in California from 2013 to 2022. Using a longitudinal panel dataset and a multilevel mixed effects modeling approach, the analysis revealed that systems with more years under private ownership were associated with a higher expected count of violations, with servicer type and primary water source modifying this effect. Explanatory models showed substantial variability at the system level and fixed effects contributed minimal explanatory power, indicating considerable gaps in relevant information. A complementary predictive analysis was performed, using a machine learning model that combined gradient boosting trees with mixed effects. Feature importances from the analysis indicated that ownership type had limited predictive value. These findings underscore the necessity for nuanced regulatory strategies to ensure water quality and regulatory compliance, irrespective of ownership type, and provide valuable insights for policymakers, water system operators, and consumers.

Table of Contents

List of Tables

List of Illustrations

List of Abbreviations and Symbols

Abbreviation/Symbol	Meaning
α	Alpha threshold for statistical significance
σ^2	Variance; residual variance
τ^2	Variance associated with random effects
ACS	American Community Survey
CA	California
CGM	Centering at the Grand Mean
CLI	Cross-Level Interaction
CWC	Centering Within Cluster
CWS	Community Water System
EPA	Environmental Protection Agency
f_i	Feature importance score
ICC	Intraclass correlation
IRR	Incidence Rate Ratio
MCL	Maximum Contaminant Level
ME-GBT	Mixed Effects Gradient Boosting Tree
MR	Monitoring and Reporting
MSE	Mean-squared error

NOAA	National Oceanic and Atmospheric Administration
NCEI	National Centers for Environmental Information
NTNCWS	Non-transient Non-community Water System
p	p-value for tests of statistical significance
PPP	Public-Private Partnership
PWS	Public Water System
R^2	Variance explained
RIFS	Random Intercept, Fixed Slope
RIRS	Random Intercept, Random Slope
RMSE	Root-mean-squared error
SLI	Same-Level Interaction
SDWA	Safe Drinking Water Act
SDWIS	Safe Drinking Water Information System
SoS	Secretary of State
TNCWS	Transient Non-community Water System
TT	Treatment Technique
UM	Unconditional Means

List of Appendices

Chapter 1: Introduction

Background

California residents receive their water through a network of systems responsible for distributing water from various sources (e.g., lakes, rivers, reservoirs, aquifers) to their taps. According to the California Department of Water Resources (2024), California government action has separated the state into regions based on natural water features and further designated over 8,000 individual public water systems (PWS) within those regions to serve the populations inside by treating, storing, and distributing the water. Management of the PWSs has changed significantly over the decades amidst rapid population growth and the resulting complexity of water resource infrastructure (Anthony, 2009). These changes have resulted in both public and private management of California PWSs and the processes that determine how water reaches its destination. The continued privatization of water in California necessitates a vigilance to ensure that the water being distributed to end users is of the same quality no matter the ownership of the PWS that distributes it – this basic right to water quality is the basis of this study on the impact of water privatization on California water regulatory compliance.

Privatization of water, in this case, will define the process by which a PWS that was previously owned, controlled, or developed by a "public" entity (local, state, or federal government) transfers ownership to a "private" entity (non-government corporation or organization). In California, water privatization has been an ongoing trend that has only increased over the decades as public municipalities have willingly transferred ownership of their water utilities to private parties (Greiner, 2016). Economic factors (e.g., recessions, tax code changes) and infrastructure challenges (e.g., maintenance, replacement of failing systems) are the

primary drivers behind the privatization of PWSs in California (Masten, 2011). The 1980's marked a significant shift towards privatization, driven largely by economic incentives when changes to the tax code encouraged private entities to invest in and take over aging PWSs owned by the government (Daly, 1993). Publicly owned water systems, especially smaller ones, often faced deterioration of PWS infrastructure that required substantial investment for maintenance and upgrades. By transferring ownership of water utility systems to private entities, public municipalities could offload costly management responsibilities while private companies could seek new investments (Rahman, 2022). The privatization trend continued into the 1990s and 2000s, with private companies identifying opportunities in the water sector. As a result, by the early 21st century, a significant portion of California's water systems had transitioned to private ownership (Ulmer, 2019).

Current State of Water Systems

As of 2024, California PWS ownership is a mix of public and private ownership. Approximately 68% of the state's water systems are privately owned (EPA's Safe Drinking Water Information System, 2024), though these systems serve a smaller proportion of the population compared to public systems which are still more prevalent in major cities. Regardless of any disparity in population, privately owned systems crucially serve millions of Californians, particularly in suburban and rural areas. Water quality standards in California are governed by both state and federal regulations, including the Safe Drinking Water Act (SDWA) (Fu, 2020). Compliance with these standards is essential for ensuring the safety and reliability of drinking water regardless of PWS ownership type. The privatization of water systems has raised questions about the ability of private entities to maintain these standards compared to their public counterparts. Studies have shown mixed results regarding the impact of privatization on water

quality on larger, nation-wide scales (Lyon, 2017). A national study by Allaire, Wu, and Lall (2018) found that while water systems in the United States generally provide reliable and high-quality drinking water, violations of health-based standards do occur. In 2015, approximately 9% of community water systems (CWSs) violated health-based water quality standards, affecting nearly 21 million people. There is also evidence for differences in cost between private and public PWSs, as studies have shown private ownership of water utilities result in 59% higher water costs on average (Bel, 2008). Proponents of privatization point to its potential to bring much-needed investment and efficiency to water systems. However, detractors raise concerns about higher costs for consumers and the prioritization of shareholder profits over public health and safety (Bel, 2020).

Problem Statement

The core issue addressed by this research is analyzing to what extent has the privatization of California water systems impacted water quality. As detractors and proponents of privatization debate the efficacy and benefits of private PWS ownership, there exists a need to determine how water quality is affected by PWS ownership while also controlling for various geographic and demographic factors specific to the location of the PWS. By including analyses on how these other factors may affect the EPA violation rate, the study can infer an impact of water privatization of California PWSs.

The significance of this study lies in its potential to inform public policy and regulatory practices and add to the pool of knowledge on water management and policy literature. By understanding how privatization affects water quality, policymakers can make more informed decisions about the management and oversight of water systems. This research aims to fill a gap in the existing literature by providing a comprehensive analysis of the relationship between water

system ownership and regulatory compliance over the past decade in California. The findings of this study have implications for various stakeholders, including policymakers, water system operators, and consumers. For public and private policymakers, the research will provide insights into the effectiveness of privatization as a strategy for managing water systems. The results of the study will also highlight potential areas for regulatory improvement to ensure that all Californians have access to safe and affordable drinking water. For water system operators, both public and private, the study will offer guidance on best practices for maintaining compliance with water quality standards. Understanding the factors that contribute to violations can help operators implement more effective management strategies. Consumers that rely on water systems may benefit from the assurance that the California water systems are being managed in a way that prioritizes their health and safety. In addition, the research will address concerns about the potential for higher costs associated with privatized water systems and the impact of these costs on affordability.

Research Questions

In order to determine the impact of privatization on water quality in California, this study seeks to determine the nature of the relationship between California PWS ownership type and adherence to water quality standards, while controlling for demographic and geographic factors that may influence regulatory compliance. The EPA Safe Drinking Water Act requires yearly mandatory reporting for California PWS, and keeps track of violations (EPA's Safe Drinking Water Information System, 2024). These violations will be used as a measure of water quality. The main questions that this study will answer are:

RQ 1. What is the relationship between California water system ownership type and adherence to water quality standards, as measured by regulatory violations of the EPA Safe Drinking Water Act?

And

RQ 2. How do county-level socioeconomic factors such as median income and poverty rate, meteorological factors such as average precipitation levels and average temperature, and political factors such as political party affiliation affect regulatory compliance for privately owned and publicly owned water systems in California?

Primary Hypothesis

The primary hypothesis of this study is that private ownership of water systems in California has a statistically significant impact on regulatory compliance with water quality standards as measured by violations of the EPA Safe Drinking Water Act. Specifically, the study hypothesizes that privately owned water systems will have different rates of violations compared to publicly owned systems, potentially influenced by differences in geographic and demographic factors discussed below.

Secondary Hypotheses

In addition to the primary hypothesis, the study will explore secondary hypotheses related to county-level factors:

- The difference in regulatory compliance between ownership types will vary depending on county-level socioeconomic, meteorological, and political factors.

By testing these hypotheses, the study aims to provide a nuanced understanding of the factors that influence water quality standards compliance for California's water systems.

Objectives

Overall Objective

The overall objective of this research is to determine the effect of private ownership on water quality standards compliance in California. By analyzing historical data on water quality violations, comparing compliance levels between public and private water systems, and assessing the influence of geographic and demographic variables, the objective will be achieved.

Specific Objectives

Analyze Historical Water Quality Data
- Collect and analyze data on water quality violations from the past decade, focusing on the frequencies of violations reported by public and private water systems as determined by the EPA.

Model Compliance Rates
- Model the degree of compliance with water quality standards between publicly and privately owned water systems over time. Modeling will involve examining different types of violations, such as Maximum Contaminant Level (MCL) violations, Treatment Technique (TT) violations, and Monitoring and Reporting (MR) violations.

Assess Geographic and Demographic Influences

    ○   Investigate how geographic and demographic factors influence water quality compliance. The investigation includes examining variations in violation rates across different regions of California and analyzing the impact of socioeconomic, meteorological, and political factors.

These objectives will guide the research methodology and data analysis, ensuring a comprehensive examination of the research question.

Limitations of the Study

Scope Limitations

The study is geographically limited to California, focusing on water systems within the state. While a one-state scope provides a detailed regional analysis, the findings may not be directly applicable to other states or regions with different regulatory environments and water management practices. The timeframe of the study is limited to the past decade (2013-2022). This period was chosen due to data availability considerations and a desire to capture recent trends. However, as a result, the study may not be reflective of longer-term historical trends or future developments.

Methodological Limitations

Data availability and quality are potential limitations of the study. The analysis relies on data from the EPA's Safe Drinking Water Information System (SDWIS) and sources from other governmental agencies like the Census Bureau and NOAA, which may have gaps or inconsistencies. Additionally, some violations may be underreported or misclassified, affecting the accuracy of the findings.

External Factors

Political and economic changes can influence water systems and their compliance with regulatory standards. Factors such as changes in government policies, economic downturns, and shifts in public attitudes towards privatization can impact the study's findings. While these factors will be considered in the analysis, their effects may not be fully accounted for.

By acknowledging these limitations, the study aims to provide a transparent and realistic assessment of the impact of privatization on water quality compliance in California.

Table 1

Definition of Key Terms

Term	Definition	Source
Privatization	The transfer of ownership and management of water systems from public entities to private companies.	Fu et. al, 2020
Water Quality Compliance	Adherence to regulatory standards for drinking water quality, as defined by the Safe Drinking Water Act (SDWA) and other relevant regulations.	California Department of Water Resources
EPA Safe Drinking Water Act (SDWA)	A federal law that sets standards for drinking water quality and oversees the states, localities, and water suppliers who implement those standards.	Environmental Protection Agency
Community Water Systems (CWSs)	Public water systems that supply water to the same population year-round.	California Department of Water Resources
Maximum Contaminant Level (MCL)	The highest level of a contaminant that is allowed in drinking water, as set by the EPA.	Environmental Protection Agency
Treatment Technique (TT)	A required process intended to reduce the level of a contaminant in drinking water.	Environmental Protection Agency
Monitoring and Reporting (MR) Violations	Failures to monitor water quality or report the results as required by regulatory standards.	Environmental Protection Agency

Summary

There is a critical need to understand the impact of privatization on water quality in California. This research aims to provide a comprehensive analysis that will inform policy decisions, improve management practices, and ultimately illuminate if privatization of water systems in California has a significant effect on water quality.

Chapter 2: Literature Review

Supplying water to citizens is a fundamental step in the building of any civilization. As the most basic and crucial of all utilities, the topics of water system technology and ownership has a long and rich body of research. Many theoretical and practical issues concerning the ownership of these water systems have been investigated, yet water quality and contamination have been largely overlooked in these analyses. The relatively recent establishment and development of California, and its rapid social and economic ascension, has resulted in an enormous collection of water supply research. However, the broader trend of disregarding water quality and contamination issues in the context of water system ownership has continued.

Early Research of Water System Ownership (through 1950)

Water systems in modern-day America date back to the Massachusetts Bay Colony (Kempe, 2006). In 1652, only 22 years after arrival, colonists in Boston installed a system capable of bringing water to both a central location (for public use) and to private homes of investors who funded the project. Boston proved to be well ahead of the curve however, as Providence, Rhode Island (1772), was the only other city with a water system in place at the beginning of the Revolutionary War. Because early water systems predate modern record keeping and data collection, information is not available on a consistent, year-to-year basis and ownership statistics are often vaguely or imprecisely reported. Numerous sources (Eutsler, 1939; Baker, 1915) report that by the year 1800, there were 16 public water supply systems in the United States, all but one of which were privately owned. This data point suggests that in-home water supplies were considered a luxury that was only available to the wealthiest households. However, supplies would change rapidly as 19th century technology advanced. Engineer and longtime water researcher M.N. Baker began producing detailed statistics in the 1880s and 1890s

as editor of the Manual of American Water Works. For the year 1890 (Baker, 1892), he recorded 2,037 water systems, of which 878 (43.1%) were publicly owned and 1,159 (56.9%) were privately owned. However, showing a pattern that has remained consistent up until the present day, the number of people served by public companies is substantially higher. For the 22,678,354 people served, the number of those using public systems was 15,018,552 (66.2%), leaving 38.8% using private water systems. These statistics show an average of 17,105 people using each public water system, while an average of 6,609 people were served by each private water system.

As previously mentioned, this paradoxical pattern has been evident throughout America's history of water supply. State level data on populations served was not included here, but for California a total of 103 water systems were recorded, with 93 (90.0%) of those being privately owned. Of particular interest to the present research, Baker (1892) includes a detailed list of every known water system that has changed ownership, either from private to public or vice versa. These include all known historical changes and include systems that were in place well before the Revolutionary War (the oldest being in Boston, established as a private system in 1652 and publicly owned since 1848). In total, 83 changes from private to public ownership were recorded, with just 17 instances of companies changing from public ownership to private. While the exact reasons for these changes are not recorded, the records of changes are still remarkably detailed and useful information. Finally, Baker noted another consistent pattern among larger cities. Among cities with a population above 20,000, the 177 total water systems were publicly owned in 99 (55.6%) of cases, and privately owned in 78 (44.4%) of cases. In a much shorter work written in 1915, Baker claims a 3:1 ratio of public to private ownership in cities with a population over 30,000 (Baker, 1915). The trend noted by the same author in 1890, that of more companies switching from private to public ownership, therefore continued well into the 20th

century. While it is possible new companies were being established under public ownership, making a change in ownership unnecessary in this context, any cities with a large population already in place would have had water systems by that point, making ownership change the more likely explanation. Even though water purity and contamination were yet to be the object of research at this early stage, Baker and others established a solid base of quantitative data that still informs an understanding of water system ownership.

A common theme of the early literature (pre-1950) on water companies is their unique place among public utilities. Most notably, at the beginning of a century that would see it become the financial capital of the world, New York City relied on private companies to provide gas, electricity, and subway service. Yet municipal (public) ownership of the water supply was so firmly established that it could be described as "the fixed and settled policy of the city of New York" (Monroe, 1906, p. 112). However, it should be noted that the world's most populous city at the time (London) was served entirely by privately owned companies, although government action, in the form of County Council regulations, was needed to raise the percentage of homes with a constant supply of water from 57.5% to 96.3% (Ashley 1906). Despite this major exception, publicly owned water supplies were typically the rule, and in "Public and Private Ownership of Water Supply Utilities," Professor Roland B. Eutsler investigated why (Eutsler, 1939). Noting that no single factor can be entirely responsible, Eutsler suggests the most common 'lay' explanation is that only municipalities can be trusted to fully comply with hygienic and sanitary practices. The author argues that while trust in compliance is certainly an important factor, it would be unwise to cite those concerns as the central factor, since hygienic and sanitary factors were not well understood until the late 19th century. While technically true, concerns about the purity of water supplies were evident long before the discovery of the

bacterium causing cholera (Koch, 1884) or the acceptance of germ theory generally (Tomes, 2000). An investigation of Glasgow, Scotland's public utilities (Crawford, 1906) reaccounts a 10-year legal battle that ended in 1855, with the municipality taking control from the private companies then supplying water. The private companies had stubbornly refused to find (a) new water source(s), despite the city's main river becoming more and more polluted over time as the city's population exploded. The Glasgow Corporation proved finding alternate sources was possible, and by 1860 the city had managed to eliminate the river as a water source entirely. This finding represents a very early example of government ownership serving the public good when for-profit companies could not, or would not, do so. The citizens knew their need for clean water was not being met, despite none of them fully understanding germ theory or water sanitation practices. A technical definition of 'water purity' is surely important, but it is not necessary for people to value what is intuitively considered to be clean water. While the research will focus on the more technical aspects of data on water contamination, it is this fundamental desire for clean water that is the basis of this project.

Eutsler's (1939) next justification for public ownership notes that such projects require such large amounts of money up front that private investors are typically either unable or unwilling to invest. In addition, government use of eminent domain is crucial to these massive projects (although Eutsler does not mention California specifically, this cruciality is especially true for the state considering that [among other factors] San Francisco and Los Angeles both have water supplies that must be transported for hundreds of miles). Eutsler also states that private water companies had difficulty in charging water rates high enough to recover their initial costs, and that these companies tended to seek other investments over time. (It is important to

note that while every argument in the preceding paragraph is plausible, and in fact highly likely, Eutsler provides no sources or data here.)

California

In *Water 4.0: The Past, Present, and Future of the World's Most Vital Resource*, David Sedlak conceptualizes the history of public water supply into four broad eras (Sedlak, 2014). The first society to import, distribute, and dispose of water on a broad scale were the ancient Romans, and Sedlak gives them sole credit for the development of what he calls Water 1.0. The early stages of drinking water treatment, and a general awareness of waterborne diseases, Sedlak labels as Water 2.0. The creative genius of Roman engineering is, of course, not the topic of this research, but is mentioned here to help place California in its historical context. Because California has become the wealthiest and most populous state, it is easy to forget how much of its development is relatively recent. In many ways, California was able to skip the many serious problems associated with Water 1.0. Granted statehood in 1850, California's early development roughly parallels the era Sedlak calls Water 2.0. The timing meant the state was able to avoid many problems historically associated with public water supplies.

One example of the kind of dilemma never faced in California occurred in June and July of 1858, when London suffered through what became known as 'The Great Stink' (Hillier & Bell, 2010). Sewage and drainage issues in London's Thames River were becoming progressively worse for decades, and the dry, hot summer of 1858 meant the stench had finally reached unbearable proportions. Similar situations arose in France (Barnes, 2006) and the Netherlands (Van Oosten, 2016), among other places. Many inland areas of California, with their arid climates and dry seasons, would have been prime breeding grounds for this extremely

unpleasant phenomenon. More importantly, California managed to avoid massive outbreaks of cholera and typhoid fever, the two deadliest waterborne diseases of the 19th century (though not completely, see Roth, 1997; Sawyer, 1916). Managing to elude these diseases was not just a matter of time, but also of place. Another distinct quality of California concerns the geography of where its cities have emerged. Throughout the history of civilization, virtually all major cities have emerged along navigable rivers, mainly due to the need for both a water supply and routes for commercial traffic. In America, at the start of the Revolutionary War, every single colonial settlement of note had been located by a navigable body of water (Standiford 2015). By contrast, all of California's earliest populations (apart from Sacramento) grew up on or near the Pacific Ocean. While those cities and their business expanded near navigable bodies of water, their locations did not satisfy their population's need for a consistent water supply. The importance of this lack of water supply cannot be overstated. The development of these cities so far from potable water sources is the single most important factor in the supply of water in California, both historically and currently, and will be a factor in virtually all the research presented here. As will be shown later, the emergence of California's most famous and populated city, Los Angeles, depended in large part on the engineering brilliance needed to overcome this geographic challenge.

Another central factor in the development of California's water systems has been the rapid growth of its population, and the corresponding need for rapid solutions with regards to water supplies. The 1850 U.S. Census, taken the same year California was granted statehood, showed 92,597 residents (U.S. Census Bureau, 1850), and while the indigenous population was not officially counted until 1880, the native population at mid-century has been estimated at roughly 100,000 (Cook, 1976). To give some idea of California's population explosion,

combining the two sources above (settler population plus native estimate) gives an 1850 state-wide total of 192,597. In 2022, this number of people would comprise just the 24th most populated *city* in the entire state (U.S. Census Bureau, 2023). Incredibly, the population of Los Angeles *doubled* from 1900 to 1904, going from 100,000 to 200,000, then to 250,000 the next year (1905). By 1915, the population had *again* doubled, reaching half a million people (Standiford, 2015). Population growth is a major consideration when accounting for what makes California unique and worthy of further study. In addition to population totals, population density is frequently associated with water system ownership, and will be explored later in this review.

California Geography

The California Department of Water Resources divides the state into ten hydrologic regions: North Coast, Sacramento River, North Lahontan, San Francisco Bay, San Joaquin River, Central Coast, Tulare Lake, South Lahontan, South Coast, and Colorado River (California Department of Water Resources, 2021). While a full description of these regions is beyond the scope of this work, any study of California's water supply requires an understanding of the major geographic regions and features involved.

The most important climatic feature of any region is the exceptionally high annual precipitation in the North Coast region. From 1989-2018, the mean annual precipitation for the state was 23.2 inches. By comparison, some areas of the North Coast region average over 100 inches of precipitation per year. Crescent City averages 70 inches of rainfall per year, a higher total than Los Angeles, San Diego, San Francisco, Long Beach, and Bakersfield combined (California Department of Water Resources, 2021). Parts of the foothills of the Sierra Nevada

located in the Sacramento River region (50+ inches per year) and the San Joaquin River region (40+ inches per year) also receive significant amounts of precipitation. The Colorado River region, in the southeast corner of the state, averages approximately 6 inches of precipitation per year. According to the California Department of Water Resources (2021), the 23.2-inch average from 1989-2018 includes a high of 83.2 inches (North Coast, 2018) and a low of less than 3 inches (South Lahontan, 2002). The precipitation in the northern area of the state is a fundamental part of California's water supply. The rain and snowmelt flowing from the higher elevations not only help to fill reservoirs and lakes, but also add significant volume to the Sacramento and San Joaquin rivers. Additionally, some of the snowmelt infiltrates the soil and eventually becomes groundwater, helping to fuel the agricultural powerhouses of the Central Valley (California Department of Water Resources, 2021).

Two additional factors central to the study of water supply are the seasonality and yearly variation in these precipitation levels. The seasonal patterns are consistent enough that the "water year" for California is defined not as a typical calendar year, but instead begins in the rainy season of October and ends the following September (OEHHA, 2022). Fully half of California's annual precipitation occurs in the months of December, January, and February. Most of the spring and summer months see very little rain, with only one quarter of the annual precipitation falling from April through September. Unlike yearly variations in total precipitation, the seasonality of California's weather has remained consistent and relatively unaffected by climate change. It should be noted here that the recent variability in yearly precipitation thought to be caused by climate change is occurring in an already variable climate. In other words, the state's precipitation levels naturally vary from year to year, so that climate change is thought to be exacerbating, rather than causing, these variations.

With regards to snow and snowmelt, one clear effect of climate change is the reduction of snowfall as a total percentage of precipitation. From 1979-2020, annual precipitation was made up of 73% rain and 27% snow (OEHHA, 2022). Historically, the highest percentages of snowfall were recorded in 1949, 1950, 1952, and 1975 (39%, 37%, 36%, and 38% respectively). Generally speaking, the ratio of rain to snow has been gradually shifting towards higher percentages of rain as total precipitation, although relatively high levels of snowfall have occurred as recently as 2001 and 2008 (both years approximately 36% of annual total. Since 2012, the percentage of rainfall has been higher than average (defined as 73%) for all but two years. The yearly balance between rain and snow is an important factor in the state's water supply. Traditionally, the "snowpack" only begins to melt as the temperature rises in mid- to late-Spring. By that point, rainfall from the higher elevations has already begun flowing into the reservoirs and rivers that make up large parts of the state's water supply. The gradual snowmelt has historically provided a second source of water later in the year, following the initial movement of earlier rainfall. Changes in the rain-to-snow ratio can result in a significant depletion of this second source of water, potentially altering allocation of water resources and making predictive modeling more difficult. Rising average temperatures constitute a serious threat to this precipitation balance in the future.

A final consideration in the study of California's water supply is the importance of the agriculture sector to the state's economy. As is so often the case, the numbers here reveal California's position at the top of United States statistical categories (California Department of Food and Agriculture 2023). In 2022, California's crop cash receipts totaled $55,871,204,000, representing over 10% of the United States total and making California the top agricultural state in the country (Iowa came in a respectable second place with over $45 billion in crop cash

receipts). Thirteen different crops from California exceeded $1 billion in value in 2022. California leads the nation in production of over seventy different crops, and produces more than 99% of America's almonds, artichokes, celery, garlic, olives, plums, and walnuts. The ten counties producing the most agricultural value are generally concentrated in the Central Valley, and the three most productive counties (Tulare, Fresno, and Kern) are all located in the Tulare Lake hydrologic region.

Water System Ownership Research Progression

Ownership Status

Unfortunately, the exhaustive research of M.N. Baker and his Manual of American Water Works represented a standard in water system research that was not matched for many decades after Baker's death. In 1974, Consumer Reports gave a rough estimate of 40,000 public water systems, with 5,500 (14%) of these categorized as privately owned (Harris & Brecher, 1974).

Water Contaminant and Disease Research

Even before water-borne diseases were well understood, it was recognized by many that the highest priority of any state regulation was the purity of the water supplied (Baker, 1915).

One breakthrough came in the early 1900s, with the discovery that simply adding chlorine to water killed off most pathogens (Sulzman 2022). By 1941, the country had over 5,000 public water systems, and 85% of those providing chlorinated drinking water. Sulzman (2022) notes that many consider this treatment to be responsible for more lives saved than any other public health advancement. With regards to specific contaminants, the three included in the present study have long been part of research on safe drinking water, as discussed below.

Arsenic

The dangers of acute arsenic exposure were known for centuries, mostly due to its notorious use as a homicidal agent (Hughes et al., 2011). Yet it was not until 1840 that the first definitive, scientific proof of arsenic poisoning was used to convict someone of murder in a court of law (Hughes et al., 2011). Since that time, deliberate arsenic poisoning has become less and less common, though there are still occasional cases. For instance, in 2003 state investigators confirmed poisoned coffee was the cause of one death and fifteen further hospitalizations in New Sweden, Maine (Zernike, 2003).

Eventually the cumulative, long-term effects of arsenic exposure became the subject of intense scientific investigation as well. A series of mid-20th century studies from Argentina showed evidence of external (skin) and internal (lung and urinary tract) cancers associated with high levels of arsenic in drinking water (Smith et al., 2002). In the 1980s, Taiwanese researchers documented an association between elevated arsenic exposure and bladder, kidney, and colon cancers (Chen et. al, 1985). The rates of bladder cancer shown were the highest ever shown to be associated with any water contaminant up to that point (Smith et al., 2002).

With regards to federal regulations, an initial interim standard for arsenic in drinking water of 50 µg As/liter was announced in 1942. In 1975, the EPA adopted the same interim standard. Not until January 2001, during the final days of the Clinton administration, did the EPA finally lower the standard to its current level of 10 µg As/liter, with a compliance deadline of January 1, 2006 (Smith et al., 2002).

Recent research has validated and continued the work on arsenic as a carcinogen. Noncancerous effects found to be associated with arsenic exposure include skin lesions,

cardiovascular disease, and diabetes (Hughes et al., 2011). Unfortunately, the presence of arsenic in drinking water remains a chronic problem in some parts of the world.

In California, routine assessments of the state's water monitoring stations record measurements of several contaminants, including arsenic. Data from 2009-2018 show the highest percentage of wells showing arsenic levels above the MCL were located in the San Joaquin River (16.3%), South Lahontan (14.7%), and North Lahontan (13.5%) regions (California Department of Water Resources, 2021).

Lead

In many ways lead is unique among the potential contaminants of drinking water included in this research project. Unlike arsenic and nitrates, there is no known benefit of lead on the human body. Historically, it was the only contaminant primarily transmitted through routes other than drinking water (currently the primary cause of lead poisoning is drinking water, only because of the successful elimination of other sources). Its presence in drinking water is a result of being transported through lead pipes, meaning testing for contamination must be done directly from the tap. Lastly, lead has been recognized as a genuine danger for millennia (Jonasson & Afshari, 2018). In fact, the earliest account of lead poisoning goes back to the second century B.C.E. Noted physician Galen (130 to 210 C.E.) noted that the lead pipes used to transport water to Rome made the water unsuitable for consumption, and later recommended that all medicines be prepared using collected rainwater instead. Pliny (23 to 79 C.E.) described the use of lead as a sweetener in wine and noted that repeated use could result in paralysis in the hands. These warnings were ignored for centuries. One researcher (Fothergill, 1812) noted continued adulteration of wine in France and Germany, and that only recently France had made adding lead

to drinking alcohol a capital offense. "On the Poison of Lead," published in the May 1812 volume of *The Belfast Monthly Magazine*, provides a helpful look at what was already known about the dangers of lead in drinking water long before California became a state. The first issue highlighted is, perhaps unsurprisingly, the use of lead pipes and cisterns to transport and store water. The explanation given here is that signs of corrosion in the pipes eventually led engineers to deduce the presence of some metal in the water. Also helpful were investigations of pipes in cases where entire families became ill because of the resultant signs of corrosion. The family members would recover after finding a safer water source. Fothergill goes on to note wine adulteration in England, referencing a popular wine-making book that recommends adding a pound of melted lead to a cask to prevent wine from turning sour. The author notes this wine-making book had been through at least six editions by that point and must have caused a great deal of lead poisoning in the process. Given all that was known about the negative health effects from lead pipe usage well over 200 years ago, the continued use of lead pipes well into the 20th century is rather curious.

Recent research on lead contamination has revealed some new areas of concern. Studies highlight that even "lead-free" plumbing can contain up to 0.2% lead, potentially leaching into water supplies (Triantafyllidou & Edwards, 2012). The corrosion of these materials, particularly in the presence of certain water chemistry conditions, exacerbates lead leaching (Masters et al., 2015). Research has consistently shown that lead is a potent neurotoxin, particularly harmful to children. Studies by Lanphear et al. (2005) demonstrate that no safe blood lead level in children has been identified, with exposure leading to cognitive deficits, behavioral issues, and various physiological problems. For adults, chronic exposure is linked to cardiovascular diseases, kidney dysfunction, and reproductive issues (Navas-Acien et al., 2007).

Nitrates

In 1945, a pediatric researcher at the University of Iowa published case reports on two infants presenting to a local hospital with a primary symptom of cyanosis, a bluish discoloration of the skin (Comly, 1945). Weeks went by without a definitive cause being identified, but eventually doctors tested the well water used to make the baby's formula and found extremely high levels of nitrate. Not long after, a second infant was admitted to the hospital with similar symptoms, and the mother reported feeding the baby with formula prepared using well water. Samples of the well water used to make formula for the two infants found nitrate levels of 140 mg/L and 90 mg/L, respectively (the current EPA Maximum Contaminant Level for nitrate is 10 mg/L). In both cases, symptoms resolved shortly after discontinuing formula made with the contaminated water. This previously unrecognized condition became known as "blue-baby syndrome," or methemoglobinemia. Soon additional reports of this condition were being published in the US, Canada (Goluboff, 1948), England (Fawns & Aldridge, 1954), and Ireland (Campbell, 1952). The potential seriousness of methemoglobinemia is underlined by a 1950 study in *Minnesota Medicine* that documented 114 cases, including 14 deaths, in the state from 1941-1949 (Rosenfield & Huston, 1950). In 1979, a nonfatal case was noted in Petaluma, California, reportedly due to contaminated well water (Fan & Steinberg, 1996). The nitrate level from the water sample was only recorded as "high," but later investigation showed extremely high levels of nitrate throughout the northwestern Petaluma Valley. It was shown that 200 wells, representing almost 40% of the area's supply, contained nitrate levels over 45 mg/L, including one well with a value of 367 mg/L. As a reminder, the EPA Maximum Contaminant Level for nitrate has been 10 mg/L for the entire history of the Safe Drinking Water Act of 1974.

One aspect of nitrates especially relevant to California's economy is agriculture, specifically the fact that agricultural runoff is the single highest source of nitrate that ends up in drinking water.

In California, nitrates must be tested either annually or quarterly, depending on the water source and history of compliance. If a laboratory finds a nitrate level that exceeds the maximum contaminant level (MCL), the water supplier must be notified within 24 hours. From that point, the supplier has 24 hours to collect and analyze a second sample (Monitoring and Compliance– Nitrate and Nitrite. 22 CCR §64432.1). As previously mentioned, state agencies routinely test the 11,178 active wells (tested at least once from 2015-2018) across the state for contaminants, including nitrate (California Department of Water Resources, 2021). Because nitrate levels are associated with agricultural runoff, the central regions of the state are expected to show the highest levels, and that is in fact the case. The percentage of wells measuring over the MCL level of 10 mg/L were highest in the Central Coast (23%) and Tulare Lake (13%) regions. Interestingly, the South Coast regions had 12% of measured wells showing a nitrate level higher than the MCL, with these results largely concentrated just north and east of Los Angeles.

In ongoing health research, a 2022 meta-analysis showed the risk of gastric cancer doubled with a 10 mg/L increase in nitrate contained in drinking water (Picetti et al., 2022). A possible association between nitrate contamination in drinking water and adverse pregnancy outcomes has recently attracted some attention, but no definitive claims have been made by health researchers to date (Royal et al., 2024; Stayner et al., 2022; Lin et al., 2023). Interestingly, increased nitrate intake in drinking water and possible effect on spontaneous preterm births has been studied in California specifically (Sherris et al., 2021)

Federal Legislation

The two landmark achievements in drinking water law remain the Clean Water Act (CWA) of 1971 and Safe Drinking Water Act (SDWA) of 1974. These two immensely important pieces of legislation, particularly SDWA, continue to underline all federal law to this day. Analysis of the Safe Drinking Water Act in the first twenty years after its passage was generally supportive, with few criticisms to be found. Writing two years after the law's passage, TJ Douglas (1976) outlined the major flaws in federal drinking water standards and how SDWA's provisions addressed these issues. The Public Health Service codified a set of Drinking Water Standards in 1914, standards which were revised in 1925, 1946, and 1962. While Douglas noted the advancements that resulted from these early policies, certain limitations were becoming more and more problematic.

State Legislation

In 1943, California created the State Water Code, which includes the following:

102. All water within the State is the property of the people of the State, but the right to the use of water may be acquired by appropriation in the manner provided by law.

104. It is hereby declared that the people of the State have a paramount interest in the use of all the water of the State and that the State shall determine what water of the State, surface and underground, can be converted to public use or controlled for public protection (California Department of Water Resources, 1957).

The state has consistently been at the forefront of legislation protecting consumers and their safety with regards to safe drinking water.

Recent Research (1980 - Present)

Global Privatization Trends

Beginning in 1990, countries throughout the world began privatizing water and sewage systems at a rapid pace (Budds & McGranahan, 2003). The trend towards privatization quickly became a global phenomenon, even though arguments for the policy shift were entirely theoretical. None of its proponents were able to reference much evidence supporting their claims, as there was no body of research showing private water systems produced better outcomes. Nevertheless, 43 countries launched water projects with some level of private participation between 1990 and 2001 (Izaguirre, 2003). Latin American and East Asia saw the highest levels of total investment, with Argentina, the Philippines, Malaysia, Chile, and Brazil being the five countries with the most private funding (Budds & McGranahan, 2003).

The outcomes of privatization in North America were similarly underwhelming. Looking back at this trend towards private sector involvement, Ohemeng and Grant (2011) analyzed the largest privatization deals in the United States (Atlanta, Georgia) and Canada (Hamilton, Ontario), respectively. In both cases, city governments terminated their contracts with private companies ahead of schedule. Soon after beginning operations, it was apparent these companies would be unable to meet their performance targets, and at times took actions that contradicted explicit promises. In Atlanta, United Water assured the city it would not lay off workers, yet just four years into their contract, early retirements initiated by the company followed by additional layoffs left the company with just 350 of its original 750 employees. The company also promised

water rates would not increase, only to raise rates in each of the four years in operation. As is typical in these works, economic considerations dominate the research, with water safety and quality only occasionally mentioned. While Ohemeng and Grant measured private operator performance by efficiency, effectiveness, cost savings, and competition, performance issues related to water quality and sanitation were also evident in both cities. Just one year into Hamilton's contract with Philip Utility Management Company (PUMC), employees had already begun reporting issues maintaining equipment along with reduced health and safety conditions. Typically, the economic focus of these papers meant these health and safety concerns were not specified or investigated in any way. Two years into the contract PUMC was responsible for a much less ambiguous incident, when system failures resulted in several million liters of untreated sewage being dumped into Hamilton Harbour. Eventually, it was the taxpayers of Hamilton and not PUMC that funded the clean-up, as well as liability claims later paid out for damages.

Federal Legislation

The 1996 reauthorization of the Safe Drinking Water Act (SDWA) marked a significant turning point in the regulation of drinking water in the United States. The legislation introduced comprehensive amendments aimed at enhancing public health protection, improving water quality, and ensuring the sustainability of drinking water supplies (Safe Drinking Water Act Amendments of 1996, S. 1316). Provisions of the new law included Consumer Confidence Reports (CCR), which requires water suppliers to provide annual reports to consumers, detailing the quality of their drinking water, including information on detected contaminants and potential health effects. This measure was intended to increase transparency and public awareness of drinking water issues. Source Water Assessment Programs (SWAP) mandated that states develop programs to assess the vulnerability of water sources to contamination. This initiative aimed to

prevent contamination at its source, thereby reducing the need for extensive treatment and lowering overall costs of operation. Perhaps most importantly, the Drinking Water State Revolving Fund (DWSRF) was established to provide financial assistance to public water systems for infrastructure improvements and compliance with drinking water standards. This fund has been crucial in supporting the modernization of aging water infrastructure across the nation. Smaller and more rural water systems often lack the financial resources to repair infrastructure, and the DWSRF addresses these gaps to help ensure overall compliance. Subsequent updates to federal drinking water laws revised rules concerning groundwater and coliform levels. A further change to federal law more relevant to the present study was the Arsenic Rule (2001), which lowered the MCL for arsenic from 50 parts per billion (ppb) to 10 ppb (USEPA, 2001). In 2021, Congress passed the Lead and Copper Rule Revisions (2021), which introduced a host of enhanced protections in the aftermath of the Flint water crisis (USEPA, 2021). However, because the compliance deadline for these provisions is October 2024, these changes will not affect any data or results in the present research.

Identified Gaps

There are several areas of potentially fruitful research concerning water systems and their ownership status. New contaminants, particularly active pharmaceutical ingredients (APIs), are emerging as potentially dangerous drinking water contaminants (Kümmerer, 2009). There are currently no comprehensive studies examining the differences in how public and private water utilities manage and mitigate emerging contaminants. Most existing research focuses on detection and health impacts, rather than on ownership models and their effectiveness in addressing these contaminants. Similarly, there is limited research on how public versus private water utilities are preparing for, and adapting to, the impacts of climate change. Studies should

focus on the resilience and flexibility of water systems under different ownership models in response to climate-related challenges such as droughts, floods, and changes in precipitation. Although installation of lead pipes was made illegal in 1986 (Safe Drinking Water Act Amendments of 1986, S. 124), many thousands of water systems contain pipes installed before that time. To date, no research has been conducted on the effectiveness of lead pipe replacements with regards to ownership status. Additionally, the recently passed Lead and Copper Rule contains provisions that will soon be a legal requirement, and investigating how different ownership models comply and manage public communication and mitigation efforts will be worthy of investigation. Finally, it is well known that smaller, private water systems often have difficulties funding the regular monitoring and testing that the EPA requires (source). It is possible that data collection technologies and software could be developed that would enhance compliance with EPA requirements and provide safer water to customers in the future.

Summary

Water delivery systems are the fundamental public utility. The body of research on the history and development of these systems, while not perfectly comprehensive, is quite large and forms the basis for the analyses in this project. The literature underscores a paradox in water system ownership: while public systems often serve more people, private ownership was more common in the early years. This trend has persisted, though the reasons for changes in ownership, from private to public or vice versa, are not well-documented. Historical examples, such as the municipal takeover of Glasgow's water supply in the mid-19th century, illustrate early instances where public ownership better served the public good by addressing issues that private companies could not or would not address. Of course, this single example illustrates just one of the many factors at play in the 19th century of public water systems. Beach's stages of

development (2022) outline how these systems were installed in progressively smaller cities over time, and these utilities were being built during a time of major advances in sanitation and technology.

The temporal and geographic contexts of California's development were significant advantages when it came to supplying its citizens with water. Many researchers (Sedlak, 2014; Beach, 2022) have marked the mid-19th century as a turning point in water infrastructure and safety. For instance, Beach specifically noted that by 1850 supply chain networks and technological innovations were emerging at a rapid pace. 1850 is also the year California was admitted as the 31st state in the Union. However, California's geography and climate, with significant annual precipitation variability and distinct hydrologic regions, presented a new set of challenges. Namely, the construction of extensive infrastructure to transport water from northern regions with high precipitation to arid southern areas.

The state's rapid population growth has further complicated water management. For instance, the population of Los Angeles doubled between 1900 and 1904 and again by 1915, necessitating rapid development and management of water resources. This growth utilized both public and private ownership models to meet the increasing demand for water, with public systems often emerging as more reliable providers.

Historical and ongoing research on specific contaminants make up a significant body of research. For example, arsenic has long been known as a dangerous contaminant, with both acute and chronic exposure linked to serious health issues. Lead contamination, primarily from old pipes, poses significant health risks, particularly to children. Nitrate contamination, largely from

agricultural runoff, has been linked to conditions such as methemoglobinemia or "blue-baby syndrome."

Federal legislation, notably the Clean Water Act of 1971 and the Safe Drinking Water Act of 1974, has been crucial in regulating drinking water quality. The 1996 reauthorization of the Safe Drinking Water Act introduced measures to enhance public health protection, improve water quality, and ensure sustainability. State legislation, such as California's State Water Code, further underscores the importance of protecting water resources and ensuring safe drinking water for all residents.

Finally, Recent trends towards privatization of water systems globally have shown mixed results, with some privatization efforts failing to meet performance targets. Current research gaps include the management of emerging contaminants, the resilience of water systems to climate change, and the effectiveness of lead pipe replacements. Addressing these gaps requires further investigation into how different ownership models handle these challenges and ensure safe drinking water for all.

Chapter 3: Methods

Data Collection

Data for this study will be collected from various governmental agencies at both the federal and state level. Water system data will be obtained from the Environmental Protection Agency's (EPA) Safe Drinking Water Information System (SDWIS) (U.S. Environmental Protection Agency, 2022). The EPA is a federal agency responsible for developing and enforcing environmental regulations, including regulations related to drinking water quality and safety (U.S. Environmental Protection Agency, 2024). The SDWIS, maintained by the EPA, contains information about public water systems and their compliance with state and national drinking water regulations. Specifically, Water System Summary (WSS) and Water System Detail (WSD) reports from the SDWIS will be used. Data will be collected for California water systems with an active status only, and only for the years 2013 to 2022. The WSS and WSD reports are quarterly data that will be aggregated annually to facilitate a year-based analysis.

WSS reports include the water system's ID number, the water system's name, the EPA region number, the primacy agency (U.S. State), the water system service type, the total population served by the water system, the cities served by the water system, the county served by the water system, the number of facilities that belong to a water system, the number of regulatory violations committed by the water system, and the number of site visits that regulators made. WSD reports contain information related to each water system's ownership type and their total number of service connections.

The numeric observations in the quarterly data appear to be running totals for the corresponding year. In other words, the numeric data either remain the same or increase across

the quarters within a particular year. In other cases, water systems may have reported data for some quarters, but not for others. To aggregate the numeric data, the maximum numeric value across the quarters will be assigned as the value for the corresponding year. For non-numeric data, the first non-missing value across the quarters will be used as the year-level value. The annual data for each year from 2013 to 2022 will be combined into one dataset, with each row representing an observation for a specific water system within a specific year, from 2013 to 2022.

In addition to EPA water system data, socioeconomic data will be collected from the U.S. Census Bureau's American Community Survey (ACS) at the county level (United States Census Bureau, n.d.). The U.S. Census Bureau is a federal agency responsible for collecting and reporting data about the nation's population and economy (United States Census Bureau, 2023-b). The ACS is an annual survey administered by the U.S. Census Bureau that provides data on social, economic, housing, and demographic characteristics (United States Census Bureau, 2024). County-level median income and poverty rate data will be obtained for California's 58 counties for each year from 2013 to 2022. At the time of this study, the ACS data for 2023 has not been released. The decision to conduct the study over the period of 2013 to 2022, rather than from 2013 to a more recent year, was influenced by this data limitation. The ACS 5-Year Estimates Subject Tables will be used rather than the 1-Year Estimates Subject Tables. The decision to use the 5-Year Estimates Subject Tables was made to ensure data consistency, as 1-Year Estimates Subject Tables are not available for the year 2020. This data may be unavailable due to challenges with data collection during the height of the COVID-19 pandemic.

Meteorological data will be collected for this study as well. This information will be obtained from the National Oceanic and Atmospheric Administration (NOAA). NOAA is a

federal agency tasked with monitoring and analyzing weather, climate, and oceanic conditions (National Oceanic and Atmospheric Administration, 2024). The specific agency under NOAA that manages its vast data archive is the National Centers for Environmental Information (NCEI) (NOAA National Centers for Environmental Information, n.d.). NCEI's County Mapping tool will be used to obtain average precipitation levels in inches and average temperature in degrees Fahrenheit for each California county from 2013 to 2022 (NOAA National Centers for Environmental Information, 2024).

Lastly, voter registration data from 2013 to 2022 will be collected from the California Secretary of State's Elections Division. The California Secretary of State is the state's chief elections officer responsible for overseeing state and federal elections and maintaining voter records (California Secretary of State, n.d.-a). This duty is facilitated by the Elections Division within the Secretary of State's Office. County-level information, such as percent of Republican registered voters, percent of Democrat registered voters, and percent of other (non-Republican and non-Democrat) registered voters, will be acquired from the Election Division's Report of Registration documents (California Secretary of State, n.d.-b). For odd-numbered years, the associated Odd-Numbered Year Report will be used. For even-numbered years, the 60-Day Report of Registration for the Statewide Direct Primary Election will be used. The 60-Day Reports will be used for even-numbered years because these reports are typically closer in temporal proximity to when the Odd-Numbered Year Reports are released. For example, many Odd-Numbered Year Reports are released in February of their respective year, and many 60-Day Reports for the Statewide Direct Primary Election are released in March or April of their respective year. In contrast, 15-Day and 154-Day Reports for the Statewide Direct Primary Election, and 15-Day and 60-Day Reports for the General Election, are more distant in time from

the month that the Odd-Numbered Year Reports are released. Using reports that are released within a similar monthly timeframe for both odd-numbered and even-numbered years is expected to maximize measurement consistency in the data.

The county-level data, including median income, poverty rate, average precipitation levels, average temperature, and voter registration demographics, will be combined with the water system data. This data collection process will yield a single dataset with water system information, such as ownership type and the number of regulatory violations committed, along with county-level socioeconomic, meteorological, and political demographic information. The dataset will be sorted by water system ID number and year, ultimately resulting in a longitudinal panel dataset (see Figure 1). This means that each row will correspond to a specific water system and a specific year, and there will be multiple observations for each water system from 2013 to 2022.

Figure 1

Data Sources and Dataset Creation

Water System Summary Report*:
- Water System ID
- Water System Name
- Servicer Type
- County Served
- Population Served
- Number of Facilities
- Number of Violations
- Number of Site Visits
- *Aggregated Quarterly to Yearly

Water System Detail Report*:
- Water System ID
- Water System Name
- Servicer Type
- Ownership Type
- Wholesaler Status
- Population Served
- Number of Service Connections
- *Aggregated Quarterly to Yearly

U.S. Census Bureau ACS:
- County Name
- Median Income
- Poverty Rate
- Yearly Data

Longitudinal Panel Dataset

CA SoS Report of Registration:
- County Name
- Total Registered Voters
- Number of Rep Voters*
- Number of Dem Voters*
- Number of Other Voters*
- *Percentages derived from data
- Yearly Data

NOAA / NCEI Mapping Tool:
- County Name
- Average Precipitation
- Average Temperature
- Yearly Data

Data Summary

The data for this study will consist of water system characteristics and county-level socioeconomic, weather, and voter registration information. Numeric variables related to water system characteristics will include: the number of regulatory violations that a water system committed, the total population served by a water system, the number of service connections that

a water system has, the number of facilities that a water system operates, and the number of regulatory site visits that regulators made.

Categorical variables will be present in the data as well. Ownership type will represent whether a water system was owned by a public entity (e.g. local, state, or federal government agency), by a private entity, or by a public-private partnership (PPP). Water system servicer type will represent the three EPA-defined classifications associated with type of population served by a water system and the system's service period. A Community Water System (CWS) is a water system that supplies water to the same population year-round and the populations are typically residential populations. A Transient Non-Community Water System (TNCWS) is a water system that does not consistently serve the same people throughout the year. Examples may include restaurants, hotels, campgrounds, gas stations, and parks. A Non-Transient Non-Community Water System (NTNCWS) is a water system that regularly supplies water to at least twenty-five of the same people for at least six months per year, but not year-round. Examples may include schools, factories, and office buildings. Each of these three water system servicer types will be represented in the variable.

A water system's primary water source will be operationalized with three distinct categories. A water system will be coded as either primarily using non-purchased or purchased groundwater, as primarily using non-purchased or purchased surface water, or as the primary water source being either unknown or a mixture of groundwater and surface water. A water system's status as a wholesale water supplier will be represented by a binary variable. Water systems that are designated as wholesale suppliers provide treated water to other water systems. Even if a water system purchases water from a wholesale supplier, that water system still has an obligation to ensure the safety and quality of the water provided to consumers. Some water

systems may be wholesalers exclusively, only providing water to other water systems, rather than to populations. These water systems may have zero values for the population served.

The data will also consist of high-dimensional categorical variables that represent a water system's unique ID number and the California county that it serves. These high-dimensional variables will be handled in a specialized manner, as detailed in later sections of this chapter. The year, ranging from 2013 to 2022, will represent the year of observation associated with a particular water system record. The expectation is that each water system will be associated with 10 yearly records from 2013 to 2022.

Numeric, county-level variables will include: median income (adjusted for inflation in the dollar value of the associated year), poverty rate (the proportion of the county population that had pre-tax money income below the official poverty threshold for that year), average precipitation in inches, average temperature in Fahrenheit, and the percentage of registered voters within a county who either registered as Republican, Democrat, or Other (i.e. not Republican and not Democrat). The rationale for variable inclusion or exclusion, and the special handling of certain variables will be described more thoroughly in later sections.

Explanatory Analysis

Quantitative Design

The primary goal of this study is to understand whether a relationship exists between water system ownership type and the number of regulatory violations that a water system commits. Given the purpose of this study, a quantitative research design was determined to be the most appropriate approach. Quantitative methods allow for the systematic examination of the relationship between variables of interest (Creswell & Creswell, 2017). A dependent variable is a

measurable outcome that may be influenced by an independent variable or multiple independent variables. The dependent variable in this study is the number of regulatory violations that a water system committed. Other variables, such as ownership type, will serve as independent variables. California water system data from the years 2013 to 2022 will be collected, combined with county-level socioeconomic, meteorological, and political demographic data from the same period, and organized into a longitudinal panel dataset with repeated measures for each water system over the period of study. This quantitative data will be analyzed using statistical techniques, which allow for meaningful conclusions to be drawn from quantifiable information (Watson, 2015).

Specifically, an explanatory analysis will be conducted via statistical modeling to determine whether a relationship exists between ownership type and regulatory compliance. Statistical models are probabilistic, mathematical approximators of the true data-generation process (Dobson, 2013). Models that adequately represent the true nature of the data can be used to make reasonable and potentially meaningful inferences about the variables of interest (Cox, 2006). In this study, multiple statistical models will be created and a final model will be used to address the research objectives. The model specification process will be based on a combination of previous research, theoretical considerations, and assumptions about the data.

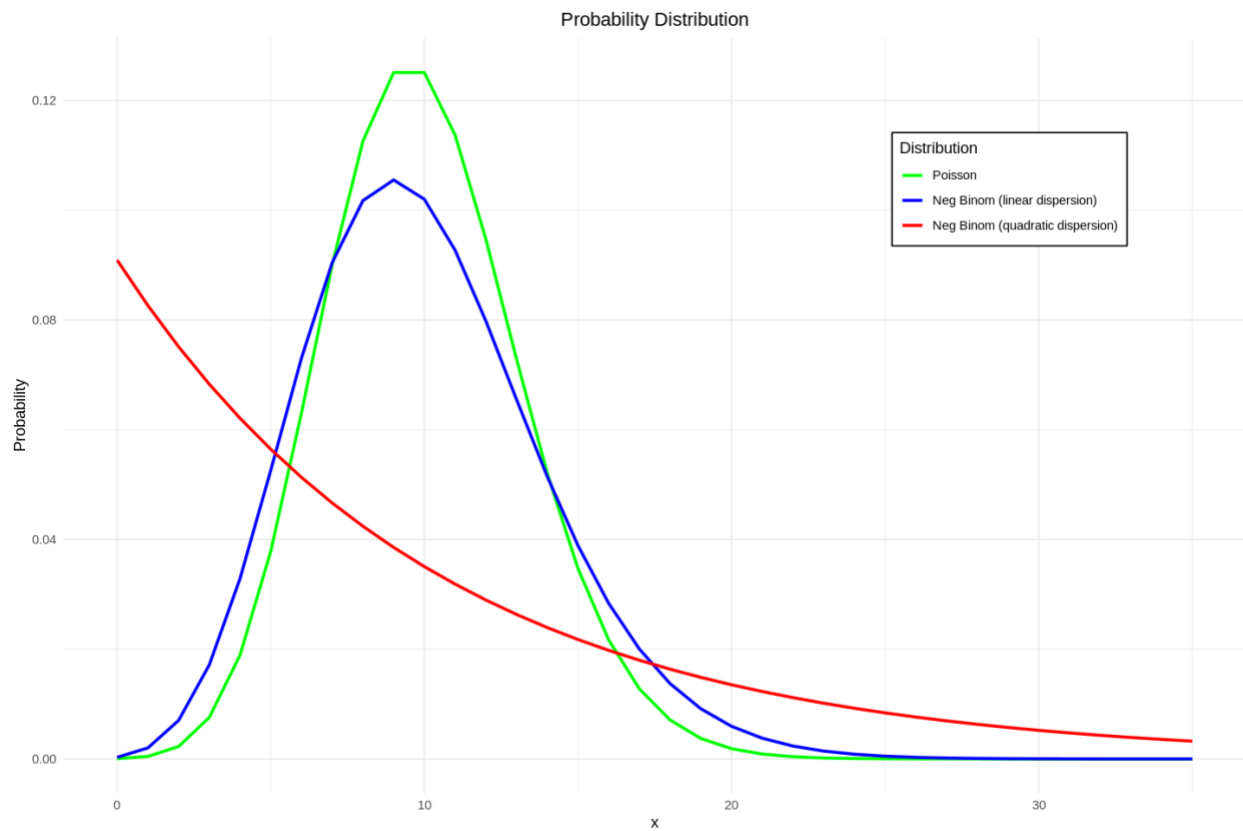*Generalized Linear Models*

Marill (2004) characterizes linear regression as a simple type of statistical model that attempts to describe the relationship between variables in terms of a straight line function. Aside from assuming a linear relationship between the dependent variable and the independent variables, linear models also assume that the model errors are normally distributed with constant

variance (Marill, 2004). When the assumptions of linear regression are violated, the model may fail to adequately represent the data, making the estimated coefficients inaccurate and unreliable (Poole & O'Farrell, 1971). Conclusions drawn from these estimates may misrepresent the truth of the underlying relationship. Standard linear regression is not appropriate for this study because many of its assumptions are violated by the data.

Firstly, the dependent variable, regulatory violations, has been shown to be highly skewed (Fu et al., 2020). Dependent variables that are not normally distributed are unlikely to have a linear relationship with the set of independent variables, thereby violating the assumption of linearity (Osborne & Waters, 2002; Williams et al., 2019). Moreover, linear models applied to skewed dependent variables are likely to produce errors that are non-normal with non-constant variance, violating other key assumptions (Bryk & Raudenbush, 1988). The rigid assumptions of linear regression are relaxed in generalized linear regression methods.

Generalized linear models (GLMs) are a class of statistical models that extend linear regression to accommodate dependent variables with non-normal distributions (McCullagh, 2019; Dobson & Barnett, 2018). GLMs require the specification of a non-normal probability distribution that more adequately aligns with the nature of the dependent variable. In addition, GLMs relate the linear combination of independent variables to the expected value of the outcome via an established link function. When the dependent variable is a skewed, non-negative count variable, such as the number of regulatory violations, a common choice of GLM is a Poisson regression model. Poisson regression uses a Poisson probability distribution and a log link function to relate the linear combination of independent variables to the link-transformed expected value of the dependent variable. These properties facilitate the improved modeling of count data (Cameron & Trivedi, 2013).

A disadvantage with Poisson models is the strong assumption made that the mean and variance of the counts are equal. This equidispersion condition is rarely satisfied in real-world data. A more common occurrence is the condition of overdispersion, where the variance of the counts exceeds the mean. If overdispersion is present in the data, the Poisson model may underestimate the standard errors and overestimate the level of statistical significance, leading to false positive results (Hilbe, 2011; Ver Hoef & Boveng, 2007). Negative binomial regression is an alternative to Poisson regression that can better handle overdispersed count data (Lawless, 1987). This method uses a negative binomial probability distribution (see Figure 2), which has an additional dispersion parameter to allow for a more flexible relationship between the mean and variance.

Figure 2

Poisson and Negative Binomial Probability Distributions

Probability Distribution

Mixed Effects Models

While the methods previously detailed may be more suitable for the data than a traditional linear regression model, there is another sticking point that must be addressed. Linear and generalized linear models operate under the assumption that observations are independent of one another (Liang & Zeger, 1993). In other words, the value of one observation does not influence or depend on the value of another observation. This assumption of independence is violated when data are arranged as repeated measures for the same subject or entity over time, as will be the case for the longitudinal data in this study. Observations from the same entity are

expected to be more correlated with one another, and more distinct from the observations of other entities, as a consequence of innate or unobserved factors (Cnaan et al., 1997). For example, an observation from Water System A may be more similar to another observation from Water System A, rather than to an observation from Water System B. This within-group homogeneity and between-group variability may be a consequence of unique system characteristics or factors not represented by the independent variables. Failing to account for the system-specific non-independence could lead to inaccurate estimates and invalid conclusions.

Mixed effects models are a class of statistical models that can handle non-independent data (Gałecki & Burzykowski, 2013). Moreover, these models can be applied within the framework of generalized linear regression for the effective modeling of non-independent count data (Bolker et al., 2009). Mixed effects models consist of two fundamental components: fixed effects and random effects. Fixed effects capture the overall mean effect of an independent variable across all entities. This component is akin to standard regression methods, which attempt to model a mean-based function. Random effects are group-specific effects that account for the group-level dependencies in the data and the group-level divergence from the mean effect (Brown, 2021).

Random effects can be specified with random intercepts, random slopes, or both (see Figure 3). Random intercepts allow each group in the grouping factor to have its own intercept as a baseline starting point. Random slopes allow the relationship between an independent variable and a dependent variable to vary across groups (Brown, 2021). By specifying both fixed and random effects, the variation in the response variable can be attributed to an overall, mean function component and a group-specific, covariance function component, allowing for more accurate effect estimations and significance tests (Oberg & Mahoney, 2007).

Figure 3

Plot of Random Intercepts and Random Slopes
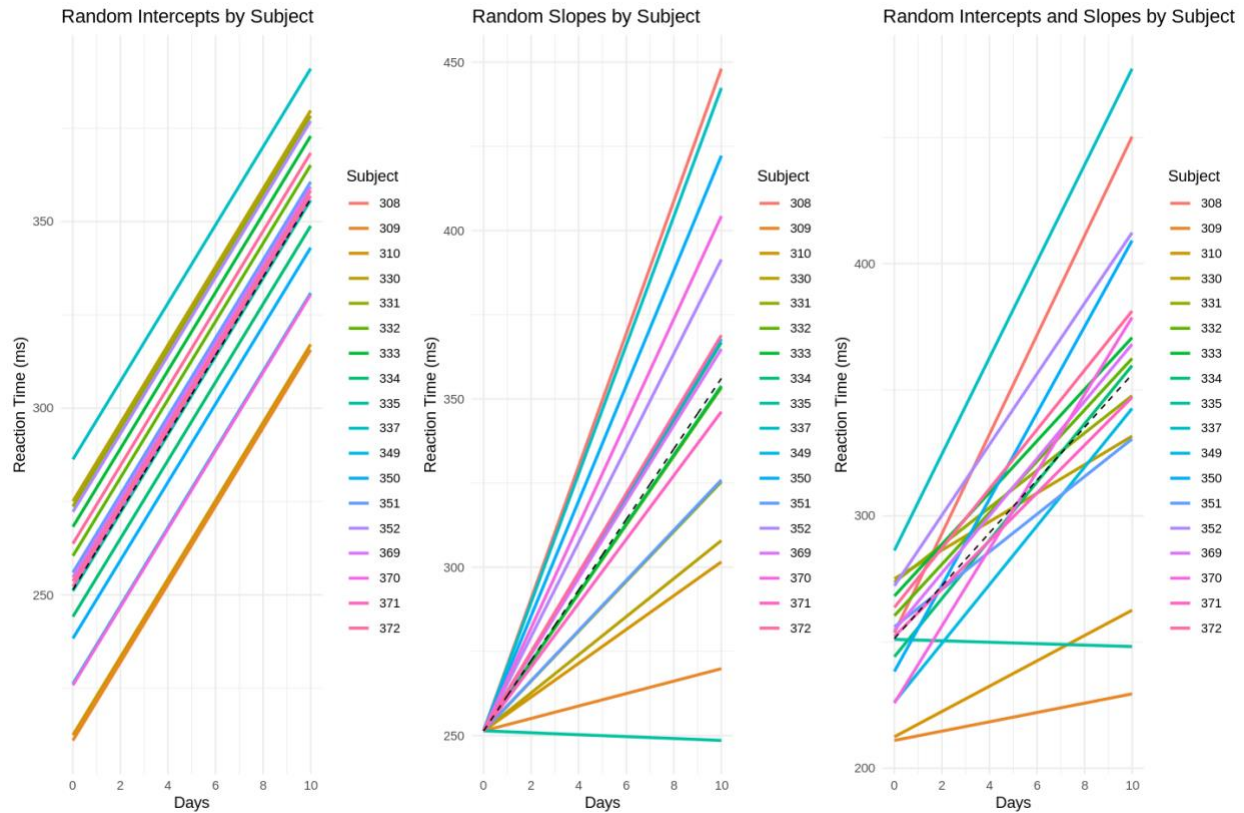
In addition to system-specific non-independence, the county that the water system serves is another potential grouping structure that could introduce non-independence. Water systems within the same county may have more similar observations than water systems from a different county. The potential similarity of observations from water systems within the same county could be due to the experience of similar social, environmental, and economic influences. When there are multiple grouping factors, the relationship between these factors can be specified in different ways. Crossed random effects occur when every level of one factor can potentially be combined with every level of another factor (Raudenbush, 1993). Nested random effects occur

when the levels of one factor are nested within the levels of another factor (Zuur et al., 2009). The data in this study are expected to more closely align with a nested effects structure, where each water system serves a specific county.

Covariance Structures

Panel data with repeated measures, as will be the case for the data used in this study, consists of both a cross-sectional component and a time component. Time is another dimension that may introduce non-independence and bias the model estimates. Observations from within the same grouping factor may exhibit correlated random effects or residual variances as a function of time (Littell et al., 2000). This temporal dependency can be accounted for by specifying a covariance structure to capture the time-dependent relationship within the same grouping factor.

Different specifications can be used for this covariance structure. One common type is the first-order autoregressive structure (AR-1) (Funatogawa & Funatogawa, 2018). The AR-1 covariance structure assumes that the highest correlation among random effects or residuals is between those from immediately adjacent time points. Moreover, the correlational dependency is expected to decay exponentially as temporal distance increases. Another common covariance structure is the compound symmetry (CS) structure (Bagiella et al., 2000). This structure assumes that the correlation between any two random effects or residuals from within the same group is the same, regardless of the time interval between them. Incorporating a covariance structure may allow mixed effects models to better capture the correlational pattern among the random effects or residuals over different time points, leading to more precise estimates and valid inference (Roy, 2006).

Figure 4

First-Order Autoregressive and Compound Symmetry Covariance Structures
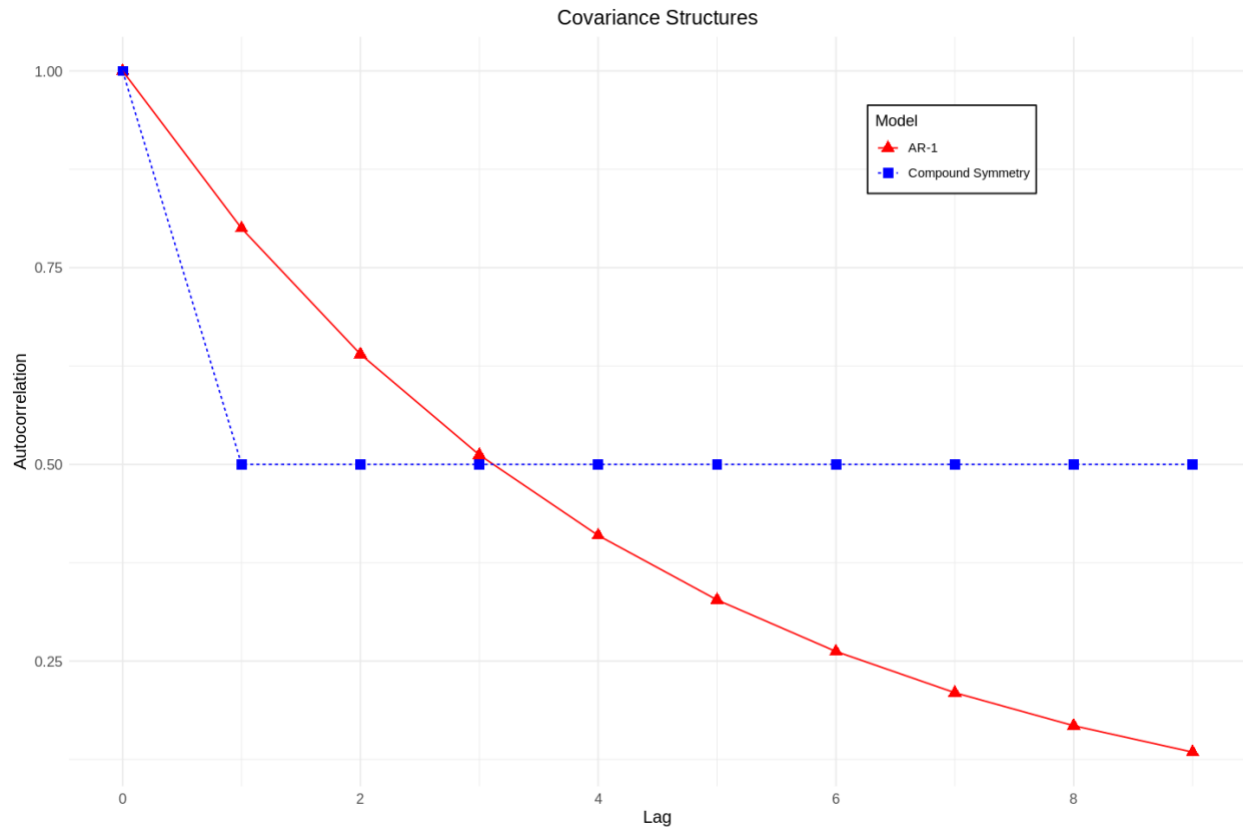
Centering in Multilevel Models

Special attention must be given to mixed effects models that incorporate a hierarchy of multiple grouping factors, as is likely to be the case in this study. Research in the field of multilevel and hierarchical modeling have shown that the effect of an independent variable on a dependent variable can be partitioned into a within-group effect and a between-group effect (Raudenbush & Bryk, 2002). Effect estimates for unpartitioned variables represent the conflated within-and-between effect of a variable, leading researchers to make incorrect and imprecise effect interpretations (Snijders & Bosker, 2012).

For example, in a two-level study with students (Level-1) nested within classes (Level-2), the effect of student socioeconomic status (SES) on academic performance can be decomposed into two distinct parts. The within-class effect measures how differences in SES among students within the same class relate to academic performance, and the between-class effect measures how differences in average SES between classes relate to academic performance. By separating these effects, researchers can more accurately identify and interpret the influences of independent variables, avoiding the pitfalls of conflated estimates that obscure true relationships and lead to erroneous conclusions (Gelman & Hill, 2007).

The decomposition of variable effects is accomplished by subtracting a variable's higher-level group mean from its lower-level values in a process known as group-mean-centering or centering-within-cluster (CWC) (Enders & Tofighi, 2007). CWC involves calculating the mean of a Level-1 variable at a higher grouping level and then subtracting this mean from the values of the original Level-1 variable. The calculated mean can then be conceived of as a new Level-2 variable that represents the contextual, between-group effect of the original variable (see Figure 5). The centered Level-1 variable now represents the isolated within-group effect. By including both the group mean and the CWC variable in the model, unconflated estimates can be obtained for the separate effects of the variable within Level-2 groups and between Level-2 groups (Hofmann & Gavin, 1998).

Figure 5

Centering-Within-Cluster Effect Decomposition

Conflated Effect

X

$X - \bar{X}_{cluster}$

$\bar{X}_{cluster}$

Within-Cluster Effect

Between-Cluster Effect
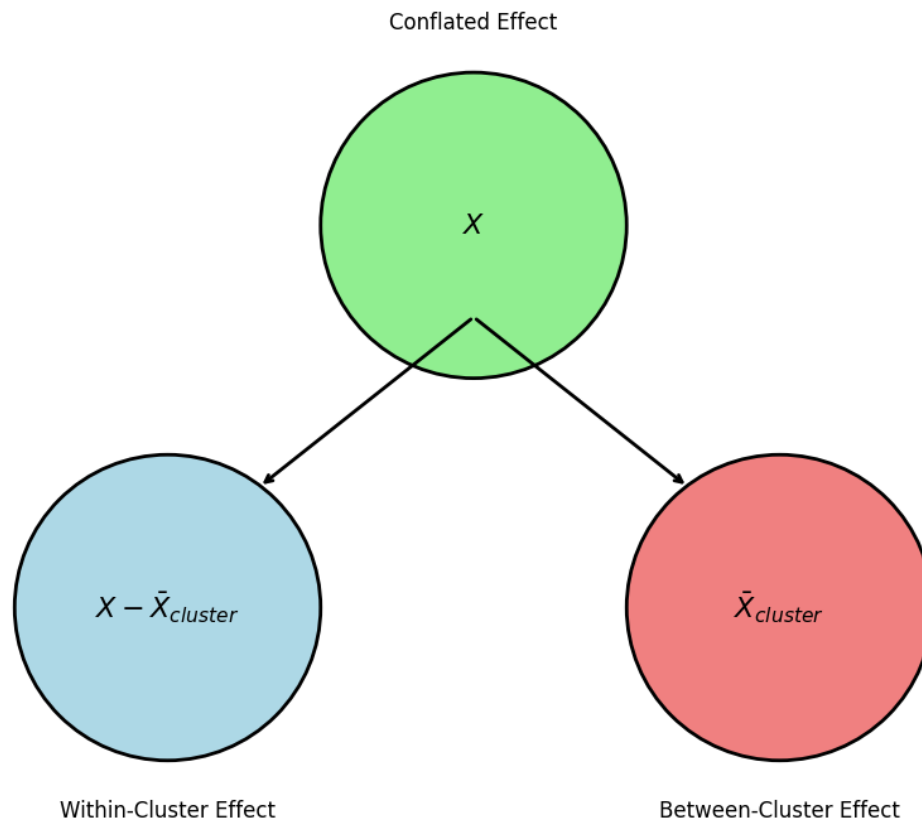
Centering also makes the model intercept more interpretable. Without centering, the intercept in a linear hierarchical model represents the expected outcome value when all covariates are zero. However, a zero value for a variable often has no meaningful interpretation. After centering, the intercept represents the expected outcome value when all covariates are at the group mean or the grand mean, depending on the centering method used. This interpretation is usually a more relevant and interesting finding to report (Kreft et al., 1995).

Centering at the grand mean (CGM) is another centering method that involves subtracting the variable's overall sample mean from its lower-level values. CGM does not disentangle within-group and between-group effects, however. The correlation between levels is still preserved. The utility of CGM is that the intercept becomes more interpretable and collinearity between main effects and interaction terms is reduced. Sometimes researchers may want to study the conflated effects as well. The highest-level variables in the model can be centered using CGM, but not with CWC since there is no higher-level grouping factor remaining (Enders & Tofighi, 2007).

The data used in this study will be arranged as a longitudinal panel dataset, with water systems and counties as potential grouping factors. This structure represents a three-level hierarchy of observations at specific time points that are nested within water systems that are further nested within counties. To obtain more valid and nuanced effect estimates, CWC centering will be implemented. Brincks et al. (2017) describe an approach for centering with three-level hierarchies referred to as CWC-1/CWC-2 (with optional CGM at Level-3). This method, hereafter referred to as CWC-1/CWC-2/CGM-3, will be adopted for this analysis.

In this method, the group means of the Level-1 variables are calculated at the Level-2 group level and the Level-3 group level. Then, a variable's initial Level-1 values are centered using the Level-2 group mean. This first step results in the decomposition of the original Level-1 variable into a within-effect component and a between-effect component. Next, Level-2 variables and the group means derived from the previous step will undergo CWC using the Level-3 group means. Brincks et al. (2017) remark that the original Level-3 variables and the Level-3 derived group means can either be left as is or be centered at the grand mean. To

promote better interpretability and computational stability, the Level-3 variables and derived group means will be centered using the grand mean.

Historically, there has been scant discussion around the centering of categorical variables. However, recent scholarship has shown that centering categorical variables is possible via CWC and CGM as well. (Yaremych et al., 2023). Centering categorical variables within clusters involves creating dummy-coded variables for all the categories within the variable, minus a selected reference category. Then, the proportion of non-zero indications (i.e. $x = 1$) at the higher group level is computed for each dummy variable. Finally, this proportion is subtracted from the dummy-coded values. The resulting transformed values represent the within-effect and the calculated proportions represent the between-effect. For CGM with categorical variables, the overall sample proportion of non-zero indications is computed and then subtracted from the dummy variables (Yaremych et al., 2023). The CWC-1/CWC-2 approach previously mentioned will also be applied to categorical variables in this analysis. The altered interpretation of these centered categorical variables will need to be accounted for too.

While many studies have promoted the avoidance of effect conflation through the centering of fixed effects, few have addressed how to avoid the conflation of random effects. Rights & Sterba (2023) showed that the decomposition of random effects in random slope specifications is crucial for obtaining unbiased estimates of within-group and between-group variability. The authors recommend including a CWC-based random slope to avoid conflation when one is interested in the effect of a Level-1 variable. If interested in a variable's Level-1 and Level-2 effects, researchers may attempt to include random slopes for both the CWC and group-mean components. Unfortunately, issues with estimation convergence may result from this more complex random slope specification.

A reasonable alternative was proposed, wherein the conflation of random effects can be avoided as long as one includes the CWC variable as a random slope and the CWC and group-mean components as fixed effects (Rights & Sterba, 2023). To avoid conflation of random effects for categorical variables, Yaremych et al. (2023) advise specifying random slopes for all CWC variables and group proportions belonging to the non-reference categories. Avoiding conflation in random effects will be attempted in this study by specifying the full within-effect and between-effect terms for the variable of interest as random slopes.

Model Specification and Inference

A common practice in multilevel modeling, especially with mixed effects models, is to carry out model specification in a progressive, step-by-step manner. This process begins by specifying unconditional means (UM) models, consisting of only the intercept term in the fixed effects component and different random intercept specifications. This initial step allows one to observe the baseline variability between the grouping factors of interest. Afterward, additional complexity is added at each subsequent stage, until a model or series of models are specified with complete fixed and random effects (Mohammadpour, 2013; Peugh, 2010; Aguinis et al., 2013; Singer, 1998).

A progressive model specification strategy will be used in this analysis. Multiple models will be fitted with increasing complexity, both at and within each stage (see Figure 6). After the UM models, random-intercept-fixed-slope (RIFS) models will be specified, adding higher-level terms for each new model. At the next stage, random-intercept-random-slope (RIRS) models will be fitted with random slope terms and covariance structures added. Following the RIRS modeling stage, new models will be specified that contain same-level interactions (SLI) between

Level-2 ownership type, representing the system-level effect rather than the observation-level

effect, and other Level-2 variables (SLI). In the last stage, cross-level interactions (CLI) will be

examined by including interaction terms between Level-2 ownership type and Level-3 county-

related variables.

Figure 6

Multi-Stage Progressive Model Specification
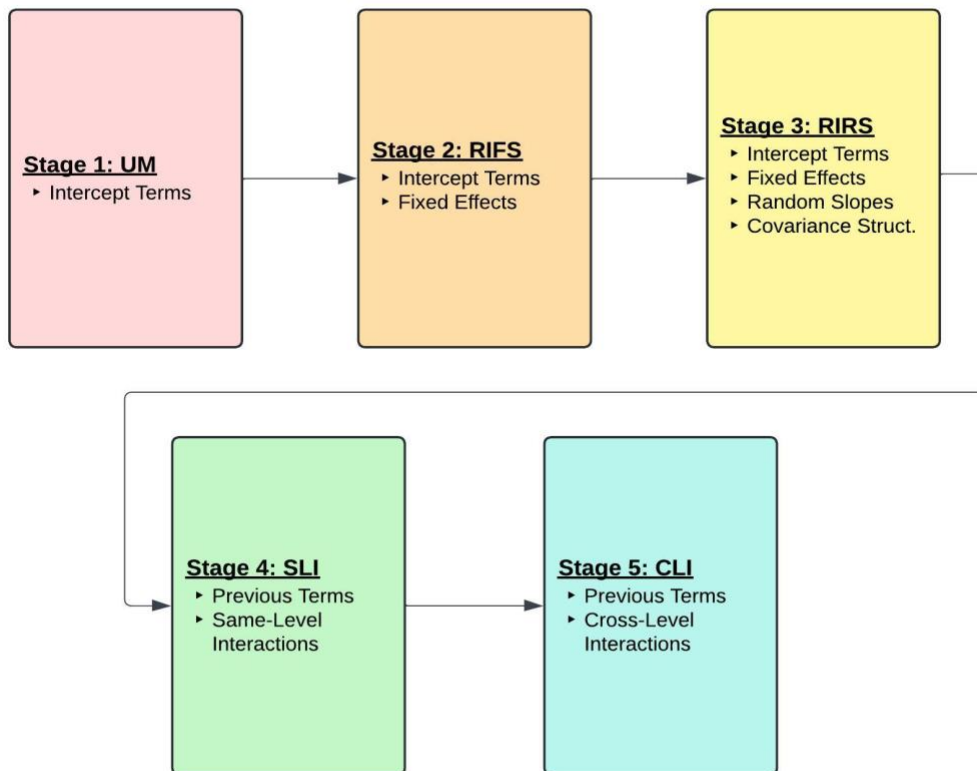
All multilevel models will be negative binomial mixed effects models to account for the

potentially overdispersed count data, the non-independence within groups, and the unobserved

sources of heterogeneity between groups. Maximum likelihood estimation (ML) will be used to

fit each model as opposed to restricted maximum likelihood estimation (REML). In general, ML

provides more unbiased estimates of fixed effects, while REML provides more unbiased estimates of random effects. However, models fitted with different fixed effects and REML cannot be directly compared because REML computations are dependent on the fixed effects design matrices (Raudenbush & Bryk, 2002). Moreover, some experts suggest that when there are a large number of groups within grouping variables and a larger number of data available, the differences in estimation between ML and REML are negligible (Snijders & Bosker, 2012). All models will be created using the glmmTMB package in the R programming language (Brooks et al., 2017).

Analysis of the specified models will consist of interpreting the fixed effects, random effects, measures of explanatory power, and residual diagnostics. Incidence rate ratios (*IRR*s) allow the interpretation of count model parameters to be on the original data's scale rather than on the log scale (Hilbe, 2011) and will be used for interpreting the fixed effect estimates. To assess explanatory power, the marginal and conditional R^2 values, as discussed by Nakagawa and Schielzeth (2013), and the intraclass correlation coefficient, will be examined. The marginal R^2 represents the proportion of variance explained by the model's fixed effects and the conditional R^2 conveys the proportion of variance explained by both the fixed and random effects. The intraclass correlation coefficient (*ICC*) is a related statistic that represents the proportion of variance explained by the random effects alone.

For residual diagnostics, scaled quantile residuals will be used rather than Pearson or deviance residuals. Dunn & Smyth (1996) observed that Pearson and deviance residuals exhibit a bias toward non-normality and heteroscedasticity when the response variable is a non-normal count variable. As such, the two proposed randomized quantile residuals as a more suitable alternative for diagnostic tests when modeling with non-normal count data (Dunn & Smyth,

1996). An extension of randomized quantile residuals involving a simulation-based approach has been studied (Correia, 2023). Scaled quantile residuals (SQRs) are based on model-simulated data, an estimated empirical density function, and a transformation of the residuals into standardized form. When a model is specified correctly, the SQRs are uniformly distributed. SQR diagnostic plots will be generated using the DHARMa package in R.

Predictive Analysis

To complement the explanatory modeling analysis, a separate modeling task will be performed to assess the predictive value of ownership type within a model trained to predict regulatory violations. In statistical inference, explanatory models use the data to closely approximate the theoretical construct space and draw valid conclusions about potentially related phenomena. In machine learning, predictive models use the data as a tool to closely approximate the underlying patterns between measured quantities and generate better predictions on unseen observations (Shmueli, 2010). A model may have high explanatory power and a variable may have a large, statistically significant effect on an outcome, but that does not necessarily mean that the model will have high predictive power or that the variable will be useful for predicting the outcome (Shmueli, 2010). Examining the role of water system ownership type within both an explanatory and a predictive context can provide a more comprehensive and well-rounded understanding of the variable and its relationship to regulatory compliance.

For this predictive task, a machine learning approach was adopted to analyze the relative importance of the ownership type variable. In supervised machine learning methods, the data is partitioned into training and testing sets, and a learning algorithm uses input features to predict a target output (Picard & Berk, 1990). While there are a wide variety of learning algorithms that

can be used for this task, decision tree algorithms naturally align with the goal of determining the relative predictive value of a feature variable (Kazemitabar et al., 2017). Tree-based models are nonparametric estimators that are highly adept at modeling nonlinear relationships and complex interactions between variables (Adler & Painsky, 2022). The models work by recursively partitioning the feature space into smaller subsets based on the most informative features. At each node of the tree, a decision rule is applied to split the data based on a specific feature and its corresponding threshold value. The goal is to create subsets that are as homogenous as possible with respect to the target variable (Adler & Painsky, 2022). Tree-based models are naturally suited for estimating feature importances by inherently selecting and prioritizing features during the tree-splitting process. The more informative a feature is for splitting decisions, the greater that feature's contribution is to the predictive outputs of the model (Adler & Painsky, 2022).

While the decision tree is a powerful estimator capable of capturing complex, nonlinear patterns in data, more advanced tree-based methods have been developed to achieve even greater predictive performance. One such development is the gradient boosting tree model, a learning algorithm that iteratively combines multiple decision trees into an ensemble (Bentéjac et al., 2020). At each step, a new tree is trained to correct the residual errors of the previous models by fitting to the negative gradient of the loss function. The predictions of all the models, including the most recent one, are then combined to produce the final predictions. LightGBM, XGBoost, and CatBoost are popular implementations of gradient boosting that possess their own unique techniques and special characteristics. However, neither implementation sufficiently addresses the challenges of modeling with non-independent, grouped data (Sigrist, 2022).

Mixed Effects Gradient Boosting

Reiterating points that were made in earlier sections, repeated measures from the same individual, entity, or other grouping factor are likely to be more correlated with one another and less correlated with the observations from other groups. These between-group differences and within-group similarities may not be explainable through the observed variables alone (Diggle et al., 2002). Mixed effects models are an effective way to handle this type of non-independent, grouped data, common in longitudinal research. In addition, mixed effects models are a more intuitive and performant tool for handling high-dimensional grouping categories (Sigrist, 2022). In recent years, machine learning algorithms have been developed that combine the advantages of mixed effects models with advanced decision tree methods such as gradient boosting tree ensembles.

GPBoost is a package that implements this type of hybrid approach (Sigrist, 2022). Using GPBoost, grouping variables can be specified to incorporate random effects in a hybrid mixed effects gradient boosting tree model (ME-GBT). GPBoost is built on top of LightGBM, so it uses an identical approach for the gradient boosting part of the model. In this predictive analysis, the Python implementation of GPBoost will be used to create a hybrid ME-GBT model. The model's fixed and random effects will be specified based on the insights gained during the explanatory modeling process. The risk of overfitting will be a primary concern in this predictive analysis, as overfitting is a common issue among tree-based models. Strategies will be implemented to mitigate this overfitting concern, including the use of a validation set with an early-stopping procedure (Zhang & Yu, 2005).

Data Partitioning and Training

The longitudinal panel dataset has both a cross-sectional element and a time element. As such, supervised machine learning can be applied to the dataset in two different ways (Smith & Fuertes, 2010). In an approach more aligned with typical cross-sectional datasets, the predictive modeling process could be designed to train a model to make predictions on unseen entities. In this case, the data would need to be partitioned in an entity-conscious way where all data for a particular water system is assigned to only one of the data partitions. Alternatively, the predictive modeling process could be designed to train a model to make predictions on unseen points in time. In this scenario, the data would need to be partitioned in a time-conscious manner where data for all water systems are assigned to one set or another based on a time-specific cutoff (Medar et al., 2017).

Since there are only ten different time points in the dataset, a time series approach may be too limited to pursue. A machine learning model would likely require data from more time points to adequately learn the underlying patterns and make accurate forecasts (Cerqueira et al., 2019). Moreover, forecasting further into the future would require knowledge of the panel data features for previously forecasted time points, which is not feasible unless additional lagged features are added and more complex modeling strategies are implemented. Given these limitations, the cross-sectional approach will be adopted instead. This approach will allow for the analysis of ownership type's importance in predicting violation counts for unseen water systems. Although this may suit the purpose of this study, it is not expected that this design will have much utility for other use cases. Perhaps there is a possibility that a model trained in this way can be used to impute missing data for water systems during the same timeframe that the model was trained on, but this speculated usage will not be explored further here.

The data will be split into training, validation, and testing partitions with 80% of the water systems and their associated observations randomly assigned to the training set, 10% randomly assigned to the validation set, and 10% randomly assigned to the testing set. This partitioning split is expected to ensure a reasonable balance between the need to provide sufficient data for the model to learn from and the need to reserve enough data for reliably testing the model's generalization ability (Sarkar, 2016). Entity-level consistency will be ensured in each partition, so that all observations from a particular water system belong to either the training, validation, or testing set. If the dataset is split in a way where observations from the same entity appear in multiple sets, then information from the validation or test sets could leak into the training process, leading to an inflated assessment of model performance (Smith & Fuertes, 2010).

The validation set will be necessary to allow for the specification of an early stopping procedure. During training, if the model fails to improve its negative log likelihood score for the set number of consecutive boosting iterations, the training process will terminate. This procedure is intended to mitigate the risk of overfitting (Zhang & Yu, 2005) and to minimize the computational burden of training the GPBoost model. Although k-fold cross-validation and nested cross-validation are more robust methods for training models and evaluating predictive generalizability (Bates et al., 2023), these methods will not be feasible with the GPBoost model and the limited computational resources available.

Figure 7

Data Partitioning Method
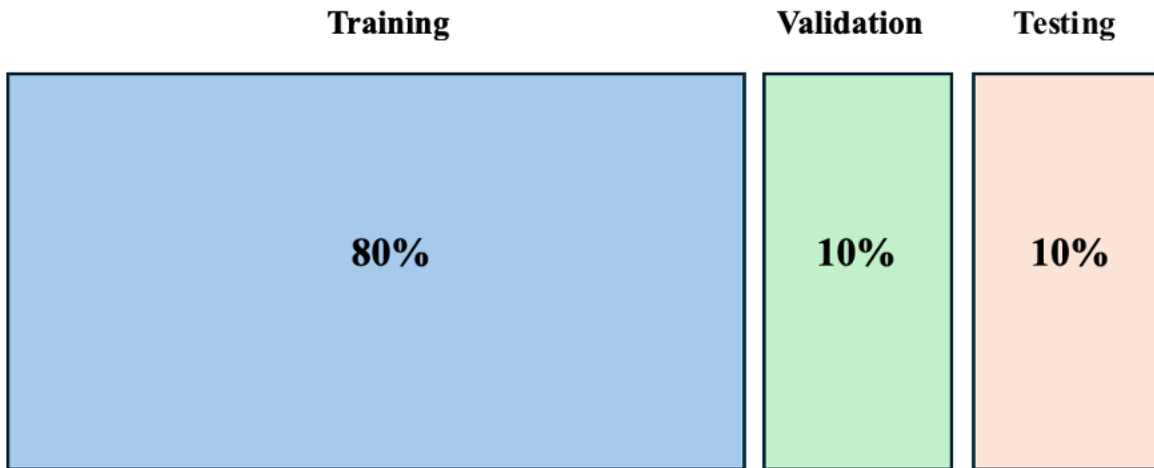
Model Evaluation

After training the model, predictive performance on the unseen test set will be evaluated by examining prediction error metrics such as mean-squared-error (*MSE*) and root-mean-squared error (*RMSE*). Once the model's predictive ability has been assessed, the feature importances of the input variables will be analyzed. Feature importances will be investigated with two different methods: split-based importance and gain-based importance. Split-based feature importances indicate the number of times that a feature is used to split the data across all boosting trees. This measure can identify which features are frequently used by the model to make decisions about splitting (Adler & Painsky, 2022). Gain-based feature importances indicate the improvement in the model's predictive ability due to the use of a particular feature for splitting. This measure may provide a more informative assessment of a feature's relative importance since it takes the predictive performance of the model into account (Adler & Painsky, 2022). Both types of feature

importance will be visually inspected and final conclusions will be made about the relative predictive value of ownership type in the model.

Summary

The primary goal of this research is to investigate whether a relationship exists between water system ownership type and EPA regulatory violations. As such, data on water system characteristics and county-level factors will be collected from the Environmental Protection Agency, the U.S. Census Bureau, the National Oceanic and Atmospheric Administration, and the Office of the California Secretary of State. A statistical approach will be implemented that uses negative binomial mixed effect models and a multilevel modeling framework. The hierarchical structure of the data and the distribution of the response variable warrant this more complex methodology. A progressive and systematic model specification approach will be adopted to conscientiously arrive at more complex, fully specified models. These models will be analyzed by interpreting fixed effects, random effects, explanatory power measures, and scaled quantile residual diagnostics. A separate predictive analysis will be conducted to determine the relative importance of ownership type as a feature within a model that predicts violation counts. This analysis will be performed using a hybrid model that combines mixed effects components with gradient boosting decision trees. The split and gain feature importances will be visualized and interpreted to ascertain the relative predictive value of water system ownership type.

Chapter 4: Results

Data Acquisition and Cleansing

Following the proposed research protocols and upon receipt of IRB approval, data on water system characteristics and violation history from 2013 to 2022 were acquired from the Environmental Protection Agency's (EPA) Safe Drinking Water Information System (SDWIS) Federal Reporting Services, specifically from the Water System Summary (WSS) reports and Water System Detail (WSD) reports. The quarterly data were aggregated at the yearly level with the strategy described in Chapter 3. Data on county-level median income and poverty rates from 2013 to 2022 were collected from the U.S. Census Bureau's American Community Survey (ACS) for all 58 California counties. County-level precipitation and temperature data for the years of interest were obtained using the County Mapping tool provided by the National Centers for Environmental Information (NCEI), an agency under the authority of the National Oceanic and Atmospheric Administration (NOAA). Voter registration demographics for each county were acquired from the California Secretary of State's Elections Division and its registration reports, using the approach previously specified. Water system and county-related data were merged by the county each water system served and the year of observation. This process resulted in the desired longitudinal panel dataset with repeated measures for each water system from 2013 to 2022.

Initially, the dataset contained observations for a total of 8,638 unique water systems. However, some water systems did not possess violation data for all 10 years in the study period. This incomplete data may be indicative of unreliable reporting practices among some water systems. Since complete data over the entire period of study was desired, potentially unreliable

water systems that did not have 10 consecutive years of non-missing violation data were excluded from the dataset. The data filtering process resulted in 6,678 remaining water systems, which was 77% of the original number of unique water systems. The 6,678 remaining water systems served 98% of the total population served by all water systems over the 10-year period. Based on the relatively high percentage of unique water systems remaining, the large number of total remaining water systems, and the population service coverage provided by these remaining water systems, the removal of incomplete water system data was not expected to significantly detract from the validity of subsequent analyses.

After filtering the data, missing values were observed for 23 records in the dataset. The county served by the water system and all county-level data were missing for these 23 records. Data missing for the county served were filled based on the observation that each water system was associated with a particular county across time. First, the earliest known value for county served was propagated forward to fill any missing values. If any missing values remained, the latest known value was propagated backward. Forward-filling first respects the temporal order of observations (Denhard et al., 2021). The most recent past value can serve as a reasonable approximation of the current year's missing data. Moreover, the missing information for the county a water system serves appeared to remain stable over time. Water systems served the same counties over the ten-year period, almost never changing. In the event that a water system was missing the second year's data, forward-filling would not be effective, since there would be no past value to propagate forward. For such cases, backward-filling based on the latest known value was deemed to be a viable alternative when values were not available to propagate forward. After this process, the missing socioeconomic, weather, and voter registration data were then filled based on the corresponding county served. After treating the missing data, there

remained one water system (CA3701893) that had missing values for the county it served. The county data was missing for all ten of its associated observations. Consequently, this water system was removed from the dataset, as there was little confidence in accurately presuming the county served based on the other data available.

Once the dataset was filtered and missing values were addressed, there were a total of 6,677 unique water systems with a complete 10 years' worth of data. This new total represented 77% of the original total number of unique water systems. Furthermore, the remaining water systems served 98% of the total population served by all water systems over the 10-year period from 2013 to 2022. The population service coverage was not altered significantly. Therefore, the removal of the single water system did not change previous conclusions about the sufficiency of the remaining data for subsequent analyses.

Variable Selection and Processing

Selection for Explanatory Analysis

The selection of variables in explanatory modeling is different from the selection of variables in predictive modeling. Unfortunately, these two types of analyses are often conflated, leading to ill-advised methodology and questionable results (Sainani, 2014). In explanatory modeling, effect estimates and statistical significance are interpreted to draw conclusions about the relationship between the exposure variable of interest and the outcome variable. The main objective is to obtain the most reliable estimate of the exposure variable's coefficient. As such, covariates are included in the model to control for factors that may affect the exposure-outcome relationship and thereby bias effect estimation. Researchers consider the causal pathway between variables and specify a limited set of hypothesized confounders, effect modifiers, and other

relevant risk factors. In predictive modeling, the primary goal is to develop a model that can most accurately predict an outcome. Covariates are included to enhance the prognostic capabilities of the model, regardless of any obfuscation of the true relationship between a predictor and the outcome. Statistical or algorithmic selection methods are often employed to narrow down a large pool of candidate variables and to mitigate overfitting. The remaining variables are those that result in the best score on a predictive metric or variables that are all statistically significant (Groenwold & Dekkers, 2023).

Figure 8

Key Differences Between Explanatory and Predictive Analyses
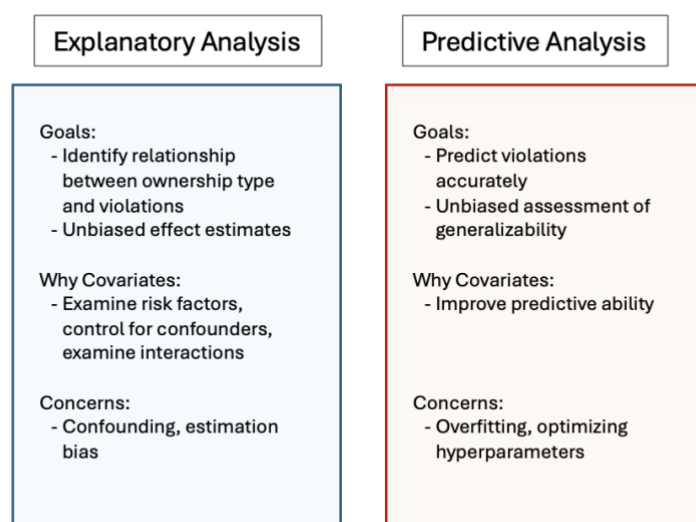
As outlined above, explanatory and predictive models have differing objectives, reasons for including covariates, and specific concerns that must be addressed. These differences should be properly accounted for rather than ignored. Additionally, researchers have shown that the

application of statistical or algorithmic variable selection to explanatory analyses can render effect estimates and measures of statistical significance too biased and unreliable for inference (Ponkilainen et al., 2021). Selecting too many irrelevant covariates or dropping statistically insignificant terms can lead to statistical significance and large effect sizes by chance alone (Muñoz & Young, 2018). A more apt approach for variable selection in explanatory modeling involves using domain knowledge, literature support, and critical reasoning to identify potential confounders, effect modifiers, and risk factors to include in the model. The causal pathway between variables and other possible sources of bias should be carefully contemplated so that researchers can obtain a reliable estimate of the exposure effect (Arif & MacNeil, 2022).

This subsection will highlight the variables selected for this explanatory analysis, including the outcome variable, the exposure variable, the grouping variables for random effect specification, and other covariates. The outcome variable in this analysis was the number of regulatory violations that a water system committed. The main exposure variable of interest was the water system's ownership type. The grouping variables were water systems, each identified by a unique identification number, and the counties that the water systems served. The year that the data was observed was included as a covariate, given the longitudinal nature of the study.

The population served by the water system, the number of service connections, and the number of facilities are all numeric measures of water system scale and service capacity. The research literature supports two different relationships regarding the influence of system scale (Rahman et al., 2010). Some researchers argue that scale may be negatively related to regulatory compliance, since larger systems can have more points of failure, where a lack of proper management yields higher contamination risk and violations. Others have proposed that system scale is associated with less violations since larger systems may have adequate resources and

infrastructure to properly maintain water quality (Acquah & Allaire, 2023). In either case, measures of scale may plausibly influence the outcome variable.

The exposure variable could also influence system scale. Private stakeholders may be less interested in operational expansion as it may entail greater costs and liability. On the other hand, system scale could influence ownership type if the transfer of ownership for larger systems from public to private stakeholders is more or less likely than the transfer from private to public stakeholders. Based on these factors, scale variables may either be confounders that affect both X and Y and should therefore be included in the model, or mediators that causally explain the effect of X on Y and should be removed to observe the direct effect of ownership type. However, system scale and service capacity may be more directly determined by factors unrelated to ownership type, so a potential mediational effect may not be an issue. Since the literature indicates that measures of system scale are important risk factors and these variables may have a potentially confounding influence, their inclusion in this analysis was determined to be warranted.

While incorporating important risk factors and controlling for potential confounders is good practice, doing so with three variables that represent the same construct could be redundant and add unnecessary complexity without further benefit. Initial exploratory analysis showed that population served, and service connections are highly correlated (r = 0.94), while population served and facilities are less correlated (r = 0.43). Population served and number of service connections may serve as similar representations of water system scale. On the other hand, the number of facilities may capture a somewhat different representation of scale. To balance adequate model specification with the avoidance of unnecessary complexity, the service connections variable was removed, and the population served and number of facilities retained.

Categorical variables that represent water system characteristics were incorporated into the explanatory analysis. EPA-designated water system type is a construct that indicates the type of population served and service consistency throughout the year. Previous research suggests that community water systems (CWS) are more likely to commit contaminant-related violations (Rahman et al., 2010). There is less clarity surrounding the relationship between servicer type and ownership type, however. Any differences in the number of privately owned versus publicly owned systems of a specific servicer type may merely stem from the association of servicer type with scale-related factors, as CWSs tend to serve more people. Because servicer type is identified as a relevant risk factor in the literature and there is interest in corroborating the results of other researchers, this information was incorporated into the modeling process.

Research related to different types of water supply sources suggests that surface water has a higher risk of contamination than groundwater (Page, 1981). This increased contamination risk could lead to a higher frequency of regulatory violations. Surface water also requires more comprehensive treatment than groundwater, which makes it more costly to use as a supply resource (Bouwer, 2000). Since private owners have an incentive to avoid costly expenses that can lower profit margins, privately owned systems may preferentially use groundwater rather than surface water. Alternatively, the type of primary water source may be more directly determined by the scale and supply demands of a water system, irrespective of ownership type. With the causal relationship unclear and the low plausibility of water source acting as a strong mediator, it was decided to include the primary water source as an additional risk factor.

The work of VanDerslice (2011) found that low socioeconomic status is associated with a higher risk of exposure to contaminated drinking water. In addition, McDonald & Jones (2018) concluded that areas with higher median income are more likely to have higher-quality drinking

water. Systems from more affluent areas may have more financial resources because of higher tax revenues or by charging higher prices to the more financially secure populace. A greater pool of resources may allow these water systems to improve or maintain water quality more easily and thereby avoid regulatory violations. In terms of potential influences on water system ownership type, socioeconomic factors may encourage or discourage private stakeholders from operating in certain areas.

If socioeconomic factors, such as county-level median income and county-level poverty rate, influence ownership type and regulatory compliance, the inclusion of these variables may control for confounding effects. However, including both variables may not be necessary since each represent a measure of socioeconomic status and were found to be highly correlated ($r = -0.77$). Median income data are highly skewed and may consist of extremely large values relative to other variables in the dataset. Even after centering the values, large numbers will remain, and can potentially negatively impact numerical stability and estimation convergence for the mixed effects model. Consequently, poverty rate was used in subsequent modeling rather than median income.

The potential influence of political demographics may warrant further consideration. Republican representatives, constituents, and media personalities tend to speak disapprovingly of government oversight, especially when related to business and environmental policy (Maibach et al., 2012). The administrators of predominantly Republican counties may possess an unfavorable attitude toward industrial and environmental regulation, resulting in water systems from these localities adopting a more lax approach to regulatory compliance. Political perspectives on regulation may also influence local officials from majority-Republican counties to be less proactive in their enforcement of regulatory compliance. Innes & Mitra (2015) found that new

Republican Congressional representatives significantly depress inspection rates for local

polluting facilities in the first year after their election. If such an effect were present in the

context of this study, then predominantly Republican counties may be associated with less robust

reporting and underestimated violation counts.

Republican politicians and constituents also possess a more favorable view of

privatization (Brooks, 2004). Predominantly Republican counties may have governing

administrations that incentivize or facilitate the privatization of water systems, leading to more

changes in ownership type from public to private stakeholders. Given the plausible association of

a more conservative political climate with lax or underreported regulatory compliance and with

private ownership, the percentage of registered voters in the county who registered Republican

was included as a possible confounder.

Other variables, such as water wholesaler status, regulatory site visits, county-level

average precipitation, and county-level average temperature were considered for inclusion in the

explanatory analysis. Wholesaler status is likely to be related to the scale and service capacity of

water systems as a consequence of the major financial and infrastructural requirements involved.

The current selection of variables already consists of plausible measures of scale. Moreover,

these variables are hypothesized to be more directly representative of the underlying construct of

scale and more informative as covariates, being non-binary terms. Therefore, the modeling

analysis excluded wholesaler status as a covariate.

The number of regulatory site visits that a water system is subjected to may be affected

by the number of violations that the water system commits. If a water system violates drinking

water regulations more frequently, then local, state, and federal regulators may be more likely to

increase their inspection frequency and level of monitoring. This increased scrutiny could pressure water systems to improve their regulatory compliance. When there is a bidirectional relationship between a potential covariate and an outcome variable, concerns about estimation bias due to reverse causality arise. If the frequency of site visits is a response to a water system's degree of regulatory non-compliance, then incorporating this information into the model could create a feedback loop where the outcome variable affects one of the explanatory variables. Roberts & Whited (2013) note that this reverse causality can lead to unreliable and inconsistent parameter estimates, making it difficult to accurately interpret effects. Given this line of reasoning, the number of site visits was omitted to prevent reverse causality issues from negatively impacting inference.

Regarding the meteorological variables, Tryland et al. (2011) and De Roos et al. (2020) found that heavy rainfall is associated with greater microbial contamination of surface water. However, research on temperature has produced conflicting findings and weak evidence for an association between temperature and water contamination (Powers et al., 2023). Because there is reasonable evidence that suggests precipitation level is a risk factor for poor water quality, county-level average precipitation was incorporated into the analysis. Conversely, researchers do not appear to have reached consensus on temperature's role. As such, county-level average temperature was excluded.

The potential modification effect that some variables may have with respect to the effect of ownership type on violation frequency was investigated. This task was accomplished by examining the estimates and statistical significance of interaction terms. Variables that were hypothesized to moderate the effect of ownership type on expected violation counts included: population served, servicer type, water source, poverty rate, and percentage of Republican

votership. In other words, the effect of ownership type on the expected count of violations was assumed to vary depending on the categorical or numerical values associated with the aforementioned variables.

Processing for Explanatory Analysis

Variables that were anticipated to change over time, such as population served and number of facilities, were preprocessed as Level-1 variables within the CWC-1/CWC-2/CGM-3 centering framework. The variables that represent water system characteristics were initially believed to be sufficiently time-invariant within systems in order to be treated as Level-2 variables. However, exploratory analysis revealed a higher degree of time-variability than expected. One or more changes in ownership type, servicer type, and water source occurred within a minority of systems over the ten-year period (3.46% of systems with at least one change, 3.94%, and 6.87%, respectively for the variables mentioned). The degree of change in system characteristics, although still relatively small, may be influential enough that failing to account for the changes may yield incorrect or misleading results. Consequently, the categorical system characteristics were preprocessed as Level-1 variables that are time-varying within systems rather than as Level-2 variables that are time-invariant or nearly time-invariant within systems. The dummy coding plus centering method from Yaremych et al. (2023) was used for these categorical variables.

The county-level variables have the same values across systems within the same county. These variables were preprocessed as Level-3 variables via CGM. The year-related variable was centered in a way that is more suitable for time variables in longitudinal studies. The first value of the year variable was set to 0, with other years following in sequential order. Centering in this

way allows the interpretation of the time variable's parameters to be more intuitive (Enders & Tofighi, 2007). In this method, the intercept represents the expected value of the outcome at the initial study year and the slope represents the expected change in the outcome as time progresses from year to year.

To facilitate a more meaningful interpretation of *IRR* estimates, the population served variable was scaled to associate changes in the expected count of violations with a 1,000-person increase in population served rather than a one-person increase. For similar reasons, the facilities variable was scaled to have estimates reflect a 10-facility increase, poverty rate was formatted as a percentage from 0 to 100 to have estimates reflect a 1% increase in poverty rate, percentage of Republican voters was formatted to have estimates reflect a 2% increase, and precipitation was scaled to have estimates reflect a 5-inch increase in average precipitation.

Selection for Predictive Analysis

As shown in Figure 9, the target variable that the ME-GBT model predicted was the number of regulatory violations. The grouping variables associated with the random effects in the model were the water systems and the county that

each served. The same variables that were selected for the explanatory analysis were also used as features for the predictive analysis, including the time-related variable. In addition, wholesaler status and county-level average temperature were included in the predictive task, despite their prior exclusion. By including more features, the ME-GBT model was expected to learn more sophisticated patterns in the data and achieve greater predictive power. Furthermore, obtaining unbiased parameter estimates is not as critical in predictive modeling. Therefore, it was not deemed necessary to consider the parameter-biasing effect of including less-relevant variables.

Lastly, a less conservative approach to feature selection was implemented in the predictive modeling task because the ultimate focus of this analysis was not to obtain the best-possible predictive performance, although that was still a secondary aim. Rather, the relative importance of each feature, and the relative importance of ownership type in particular, was the true, desired end product. Including more features may allow for conclusions to be drawn about ownership type's predictive importance relative to a wider range of variables. In other words, it is of interest to know which features the tree-based model considers unimportant for prediction. Prematurely removing features may lead to a less comprehensive understanding of relative importances, with some noted exceptions.

Indeed, there were certain features that were omitted from the predictive analysis when their removal was expected to improve analytical clarity. For example, research suggests that tree-based models have a tendency to assign high importance to highly correlated features (Strobl, 2008). Since service connections were highly correlated with population served ($r = 0.93$), and median income was highly correlated with poverty rate ($r = -0.77$), these variables were excluded to avoid obtaining biased importance scores. Regulatory site visits was another variable that was excluded from the predictive model, due to concerns about target leakage. When a feature variable inadvertently reveals information about the target variable, this information leakage can lead to overly optimistic prediction results (Rosenblatt, 2024). As hypothesized previously, the number of regulatory violations may influence the number of times that regulators visit a water system's sites. If this assumption is true and a water system has a higher number of site visits, this information may unintentionally signal to the model that the water system must also have a higher number of regulatory violations. To avoid data leakage and misleading results, the number of site visits was omitted.

Figure 9

Predictor Diagram for Mixed Effects Gradient Boosting Tree Model
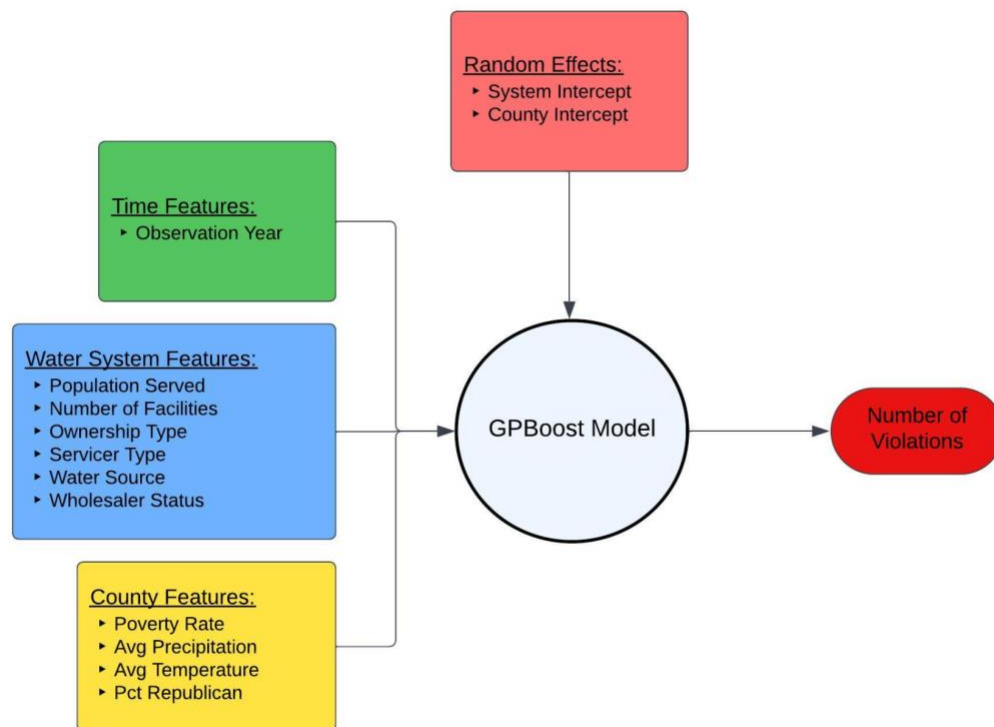
Processing for Predictive Analysis

The CWC-1/CWC-2/CGM-3 method of preparing variables for multilevel modeling was used again to preprocess the features for the predictive model. Including the decomposed, level-specific features would allow for a more nuanced analysis regarding which features were important and at which level of the hierarchical structure the features were important. The year variable was centered in the same way as the previous analysis. The categorical features were dummy coded and also adhered to the CWC-1/CWC-2/CGM-3 approach. The additional steps

taken to make estimates more interpretable in the explanatory analysis were not applied when processing features for the predictive analysis, as coefficient or *IRR* interpretation was not the aim here. In sum, the major changes in variable specification from the explanatory model to the predictive model were the inclusion of two additional variables, wholesaler status and county-level average temperature. The previously implemented preprocessing strategy was applied to these new features, without additional steps to aid in effect interpretation.

Descriptive Statistics

The dataset used in this study included 66,770 data points from 6,677 unique PWSs in California. Table 1 shows the summary statistics of the key variables measured in the dataset.

Table 2

Summary Statistics of Key Variables

Variable	Mean	Min	Median	Max	Std Dev
Violations	9.79	0	5	278	13.11
Population Served	6,363.81	0	125	4,085,000	65,133.90
Service Connections	1,455.63	0	16	727,143	12,534.45
Facilities	8.95	0	4	1,784	23.31
Site Visits	12.15	0	8	287	14.06
Median Income (County)	$64,090	$34,974	$60,704	$153,792	$17,748
Poverty Rate (County)	0.16	0.049	0.152	0.283	0.05
Avg Precipitation (County)	21.13	2.06	16.12	97.59	16.09
Avg Temperature (County)	60.09	43.5	60.2	76.5	5.08
% Republican (County)	0.32	0.064	0.332	0.552	0.09

Note. Based on a count of 66,770 observations for each variable.

The mean number of violations across all observations was 9.79, with a standard deviation of 13.11. The standard deviation being relatively higher than the mean indicated a significant degree of variability in compliance among the different water systems. The maximum number of violations recorded was 278, highlighting the extreme non-compliance of certain systems. The presence of water systems with zero violations suggested that some systems were

able to maintain full compliance with water quality standards. The population served by water systems varied widely, with a mean of 6,363.81 and a standard deviation of 65,133.90. The median population served was 125 people, which was considerably lower than the mean. This behavior is consistent with a heavily right-skewed distribution (see Figure 10), where a few systems serve very large populations compared to the majority of water systems. The mean county-level median income was $64,089.07, with a standard deviation of $17,748.36. This variation in income levels across counties could influence water system compliance, as wealthier counties may have more resources to invest in water infrastructure. The mean county-level poverty rate was 0.16 (16%), with a standard deviation of 0.05 (5%). The range from 0.049 to 0.283 indicates that some counties had significantly higher poverty rates, which may be associated with higher violation rates due to resource constraints. The mean for average county-level precipitation was 21.13 inches, with a standard deviation of 16.09 inches. The variability in precipitation could affect water quality and system performance, as different regions face distinct environmental challenges. The mean for county-level average temperature was 60.09°F, with a standard deviation of 5.08°F. Variations in temperature can impact water quality, particularly in terms of biological activity and contamination risks. The mean percentage of Republican affiliation was 0.32, with a standard deviation of 0.09. Political dynamics at the county level may influence water policy and regulatory enforcement.

Figure 10

Overall Distribution of Regulatory Violations

Histogram of Violations
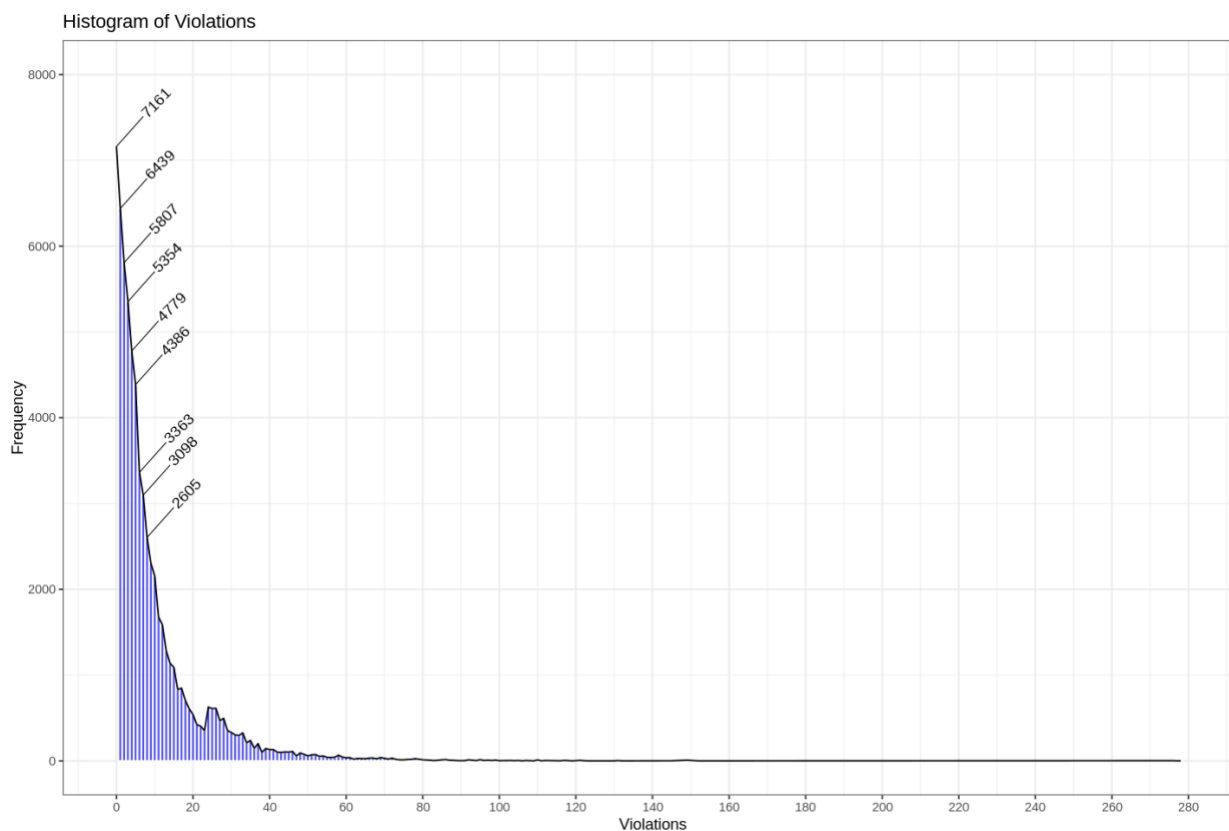
Differences in regulatory compliance by ownership type were observed prior to the

modeling analysis. The proportion of water systems present in the dataset were as follows:

- Private: 43,493 systems (65.14%)
- Public: 22,425 systems (33.59%)
- Public-Private Partnerships (PPP): 852 systems (1.28%)

The following table includes summary statistics for the key variables, delineated by ownership

type.

Table 3

Summary Statistics for Key Variables by Ownership Type

Variable	Ownership Type	Count	Mean	Min	Median	Max	Std Dev
Violations	Public	22,425	8.28	0	4	278	13.33
	Private	43,493	10.57	0	6	145	12.93
	PPP	852	10.09	25	6	96	13.14
Population Served	Public	22,425	15,530.43	0	370	4,085,000	108,451.10
	Private	43,493	1,757.86	0	100	1,007,514	19,583.73
	PPP	852	219.37	1	100	2,800	462.55
Service Connections	Public	22,425	3,518.63	0	24	727,143	20,580.49
	Private	43,493	419.97	0	15	225,800	4,417.12
	PPP	852	25.61	0	5	501	54.80
Facilities	Public	22,425	13.42	0	5	1,784	34.60
	Private	43,493	6.62	0	4	786	14.15
	PPP	852	9.83	0	10	51	7.30
Site Visits	Public	22,425	13.19	0	9	287	15.91
	Private	43,493	11.70	0	8	202	13.09
	PPP	852	7.63	25	6	51	5.15
Median Income (County)	Public	22,425	$63,279	$34,974	$58,861	$153,792	$18,340
	Private	43,493	$64,676	$34,974	$61,276	$153,792	$17,448
	PPP	852	$55,447	$41,924	$51,261	$124,055	$13,550
Poverty Rate (County)	Public	22,425	0.16	0.05	0.16	0.28	0.05
	Private	43,493	0.16	0.05	0.15	0.28	0.05
	PPP	852	0.20	0.07	0.21	0.28	0.05
Avg Precipitation (County)	Public	22,425	21.26	2.06	16.04	97.59	16.59
	Private	43,493	21.05	2.06	16.14	97.59	15.90
	PPP	852	21.92	3.12	20.72	75.84	12.31
Avg Temperature (County)	Public	22,425	59.77	43.50	60.00	76.50	5.61
	Private	43,493	76.50	43.50	60.40	76.50	4.81
	PPP	852	58.70	47.90	57.90	57.90	3.26
% Republican (County)	Public	22,425	33	6	34	55	9
	Private	43,493	32	11	32	55	9
	PPP	852	38	11	40	47	7

A difference was observed in average violation counts between ownership types, as public systems had a mean of 8.28 violations per year, while systems with private ownership had a mean of 10.57 violations per year and systems with PPP ownership had a mean of 10.09 violations per year. The higher violation average for private and PPP systems suggests that these

privately and PPP owned systems may face more challenges related to maintaining regulatory compliance. Public systems served a mean population of 15,530.43, with a median of 370 and a standard deviation of 108,451.10. Private systems served a smaller mean population of 1,757.86, with a median of 100 and a standard deviation of 19,583.73. Public systems tend to serve larger populations compared to private and PPP systems. This difference in scale could influence the resources available for maintaining compliance and the complexity of operations. Data was also aggregated at the county level for the water systems. Figure 11 illustrates that from 2013 to 2022, counties located in the central areas of California experienced a higher average number of violations. The county-level data indicate the presence of structural factors that may increase the risk of regulatory non-compliance for water systems in these central counties. Another possible explanation is the lower population density in central California. If the lower population density is linked to there being fewer water systems in these counties, then a small number of systems with high violation counts could disproportionately elevate the average violation count.

Figure 11

Map of Mean Violations by County
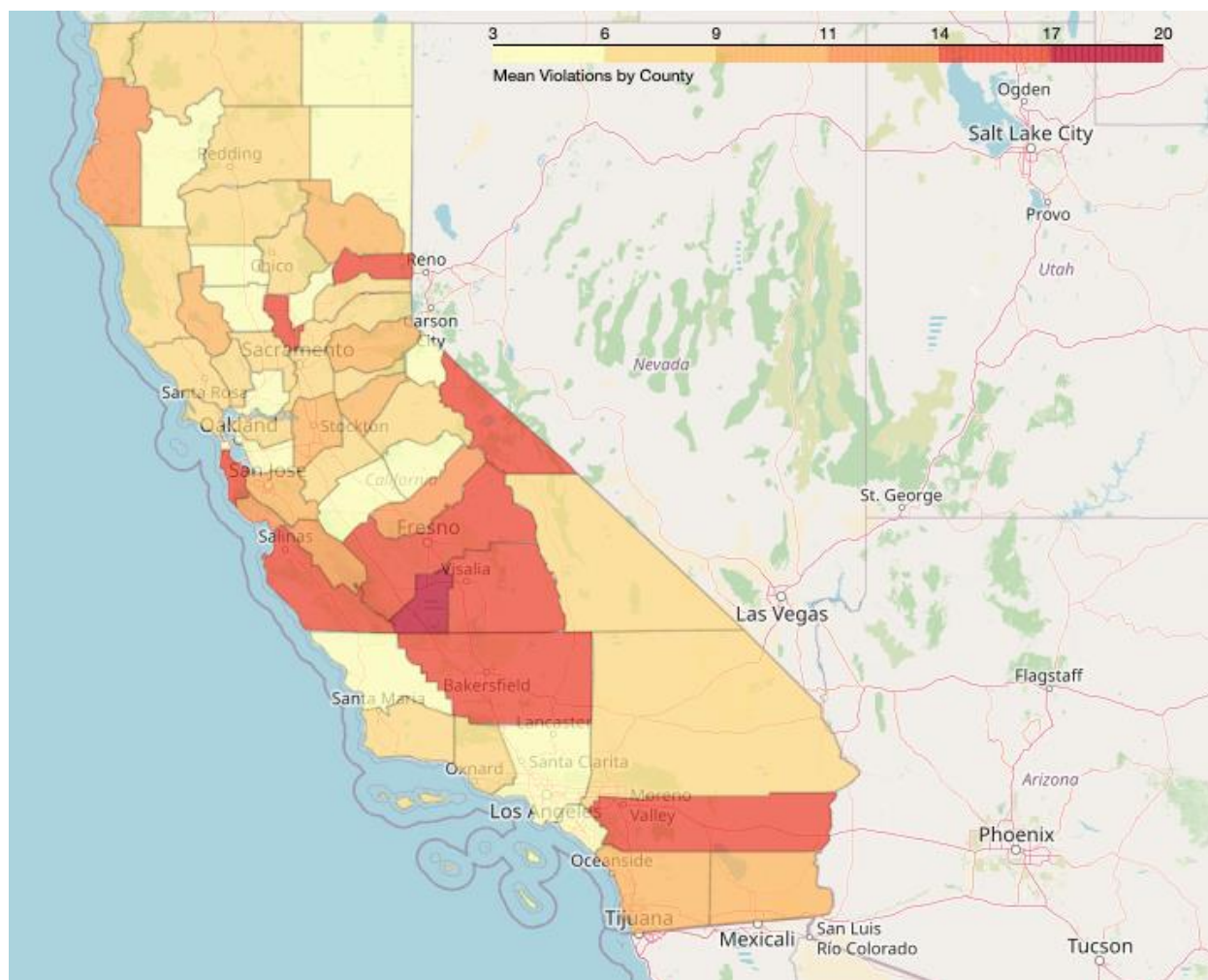
Explanatory Modeling Results

A series of multilevel negative binomial mixed effects models were fitted to examine the relationship between water system ownership type and regulatory violations. Relationships with regulatory compliance were also examined for other potential risk factors. Centering techniques were used to decompose within-effects and between-effects. A five-stage process was implemented to systematically analyze increasingly complex model specifications. This

progression enabled a more nuanced and comprehensive analysis of fixed and random effects and their stability across different model structures. Only fixed effects that were statistically significant at $\alpha = 0.05$ will be discussed at length.

Stage 1: Unconditional Means (UM)

Unconditional means (UM) models contain only an intercept term in the fixed component of the model and different grouping-specific intercept specifications in the random component. Examining the UM models can validate whether multilevel modeling is warranted and provide insight into the distribution of total variance across different levels (Goldstein, 2011; Park et al., 2022). In this stage, three separate UM models were fitted (see Figure 5). The first model (UM-1) contained no random intercept terms, the second model (UM-2) contained only a system-specific random intercept, and the third model (UM-3) contained a random intercept term representing water systems nested within counties.

Figure 12

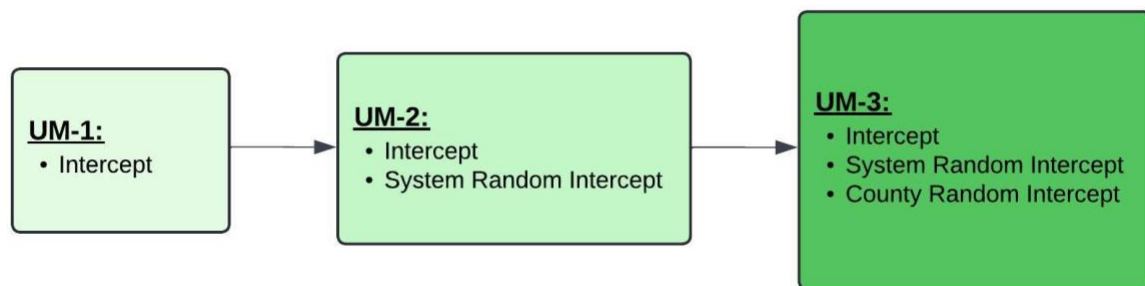*Unconditional Means Model Progression*

Table 3 shows the results of the UM models. The first model without a random intercept term had a statistically significant *IRR* of 9.79, representing the overall expected count of regulatory violations across all observations, without considering any clustering or variation between water systems and counties. The *IRR* for UM-2, which had a random intercept for water systems, was statistically significant with a value of 4.43. A discrepancy in *IRR*s of 5.36 suggested that ignoring the underlying hierarchical nature of the data may introduce significant bias. After accounting for the unobserved heterogeneity at the system level, the remaining variation within systems was low, as evidenced by the residual variance of 0.21. However, variation across systems was greater ($\tau^2 = 2.25$).

An *ICC* of 0.91, indicated that the system-specific random intercept explained approximately 91% of the variability between systems. The model with a nested grouping structure of systems within counties had a statistically significant *IRR* of 3.88, lower than the previous UM models. However, the uncertainty surrounding the third model's estimates was greater (*SE* = 0.04 for UM-1, 0.08 for UM-2, and 0.31 for UM-3). The residual variance and *ICC* measures did not differ substantially from the second model ($\sigma^2 = 0.24$ and *ICC* = 0.90 for UM-3). Random effect variances revealed that some of the variability across systems can be explained by county-level differences, as accounting for the county grouping structure reduced variation at the system level ($\tau^2 = 2.25$ for UM-2 to $\tau^2 = 1.92$ for UM-3). The analysis of the UM models suggested that a multilevel modeling approach was indeed necessary to account for the underlying structure of the data and to obtain more reliable estimates.

Table 3

Unconditional Means Models

	Model 1	Model 2	Model 3
Intercept	9.79*** (0.04)	4.43*** (0.08)	3.88*** (0.31)
Variance components			
Observation-level variance		0.21	0.24
System-level variance		2.25	1.92
County-level variance			0.33
Variance explained			
Intraclass correlation		0.91	0.90
Marginal / Conditional R^2		0.00 / 0.91	0.00 / 0.91

Note. Total $N = 66,770$; ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Estimates and standard errors are presented as Incidence Rate Ratios.

Stage 2: Random Intercept, Fixed Slope (RIFS)

The second stage of model-building involved the progressive addition of fixed effects. All models were defined with a nested random intercept term for systems within counties. The CWC-1 model included the within-system effects of the Level-1 variables that varied within systems, plus the centered time variable. The CWC-1/CGM-2 model consisted of the same fixed effects as the previous model, in addition to the grand-mean-centered between-system effects derived from the Level-1 variables. A third model, processed via CWC-1/CWC-2, centered the derived between-system effects within the county level rather than at the grand mean. The last model in this stage built upon the CWC-1/CWC-2 model by incorporating county variables and derived Level-3 terms that were grand-mean-centered (CWC-1/CWC-2/CGM-3). Since all counties within the state of California were included in the analysis, the grand mean was interpreted as the state-level mean to make interpretations more contextually relatable. The

similarities and differences between the models will be analyzed next, starting with the fixed effects.

Models at this stage had statistically significant effects for year, with *IRR*s remaining consistent at 1.02 for most models, except for the three-level model, which had a slightly lower *IRR* of 1.01. These estimates suggested that, on average and holding all other variables constant, a 1 or 2% increase in the expected count of regulatory violations was anticipated from one year to the next. In other words, if a water system had 100 violations in 2013, increasing to 101 or 102 violations in 2014 would be expected, on average. However, a large number of systems had violation counts well under 100 and closer to 0 or 1, in which case, a 1 or 2% increase does not have much practical relevance. A violation is either committed or not; there cannot be a 1% violation. Ultimately, given the small effect size, changes in violation counts over time may not be substantial within systems.

For all four models, a ten-facility increase above the system mean was associated with, on average, a statistically significant 1% reduction in the expected count of regulatory violations (*IRR* = 0.99 for all models; p = 0.35, 0.34, 0.34, and 0.17, respectively). Within a given water system, observation years where the system was a TNCWS servicer type were associated with a 6 or 7% average decrease in violations compared to years where the system was a CWS servicer type, holding all other factors constant (*IRR* = 0.94, 0.94, 0.94, 0.93, respectively; $p < 0.01$ for all models). However, systems do not frequently change servicer type over time. Consequently, a small effect associated with changing servicer type may have little practical significance. Compared to groundwater usage, surface water usage during an observation year was associated with a 7% increase, on average, in the expected count of violations within a given water system,

holding all other factors constant ($IRR = 1.07$; $p < 0.01$). This finding was consistent for all four models, regardless of specification complexity.

Next, the Level-2 fixed effects from the CWC-1/CGM-2 model were analyzed and compared with the Level-2 fixed effects from the CWC-1/CWC-2 and CWC-1/CWC-2/CGM-3 models. The analysis showed that grand-mean-centering the Level-2 variables, which in theory preserves the correlation between the county-level and system-level effects, did not lead to notably different estimates compared to centering-within-cluster, which is expected to break up this effect-conflating correlation. Given that the fixed effects did not contribute much in terms of variance explained (Marginal $R^2 = 0.08$ at most, across the models), it may be the case that the effects are not strong enough to observe sizable differences in effect partitioning methods. Alternatively, county-level and state-level aggregates may be similar enough that centering by either tends to yield similar results. Since CGM-2 estimates were overwhelmingly similar to CWC-2 estimates, the CWC-2 interpretations will be provided for the remainder of this section.

A water system that had an average number of facilities ten units greater than the county average was associated with a 2% decrease in the expected count of violations across the multilevel models ($IRR = 0.98$; $p = 0.03$). Therefore, the magnitude of the facilities-related effect was relatively consistent from the within-system level to the between-system level. Within a given county, water systems with exclusively private ownership over the ten-year period were associated with a 60% increase, on average, in expected regulatory violations, compared to water systems with exclusively public ownership ($IRR = 1.60$ and $p < 0.01$ for all multilevel models). Put another way, an additional year of private ownership rather than public ownership was associated with a 6% increase in the expected count of violations. This result suggests that a

greater number of years spent under private ownership is a significant risk factor for regulatory non-compliance.

For all two-level and three-level models, water systems within a given county that were classified as either NTNCWS or TNCWS for the entire study period had, on average, a 34% decrease ($IRR = 0.66$, $p < 0.01$) and 57% decrease ($IRR = 0.43$; $p < 0.01$), respectively, in expected violations. This outcome is consistent with the research literature on this topic, which has suggested that CWS-classified systems tend to commit more violations (Kirchhoff et al., 2019). Within a given county, water systems that exclusively used surface water had roughly 17% less violations, on average and holding all other factors constant, compared to water systems that exclusively used groundwater ($IRR = 0.83$ for the multilevel models; $p < 0.01$ for all). This finding appears to be a departure from the association found at the within-system level.

The Level-3 fixed effects from the CWC-1/CWC-2/CGM-3 model specification will be addressed next. An IRR of 0.99 ($p = 0.03$) for the county-level population served variable indicated that counties possessing an average service population one-thousand-people higher than the state mean, had no practically significant difference in violation counts. Additionally, counties with a one-percent higher poverty rate, relative to the state mean poverty rate, were expected to have no difference in the count of violations, holding all other variables constant ($IRR = 1.00$; $p < 0.001$). Lastly, counties with a two-percent higher percentage of Republican registered voters, compared to the mean percentage at the state level, were associated with a 2% lower count of violations, all other variables being held constant ($IRR = 0.98$; $p < 0.001$).

In terms of the random effects, residual variances remained the same across all the models ($\sigma^2 = 0.23$), indicating low overall variability in violation counts within systems. *ICC*

values were roughly identical to the UM models (ICC = 0.91, 0.90, 0.90, and 0.90, respectively).

The CWC-1 model that contained the within-system effects of system-related variables showed

higher variation between systems than the two and three-level models (τ^2 = 1.92, 1.74, 1.74, and

1.74, respectively). Across-county variation remained relatively stable between the different

models, but was lower for the model where county-level covariates were included (τ^2 = 0.33,

0.32, 0.33, and 0.29, respectively). The marginal R^2 values changed from nearly zero in the first

model, to approximately 0.07 in the two-level models, and finally to roughly 0.08 in the most

complex model. Adding within-system fixed effects had virtually no impact on explanatory

power. While the inclusion of Level-2 and Level-3 terms improved explanatory power, the total

proportion of variance explained by the fixed effects remained extremely low, especially

compared to the variance explained by the random effects. Such severe, group-level

heterogeneity and the inability of the fixed effects to account for much variability may suggest a

lack of sufficiently relevant variables to account for system differences in regulatory compliance.

Table 4

Regression Results for the 4th RIFS Model

|	Estimate	SE	CI	p
Intercept	3.76	0.31	3.21 – 4.42	<0.001
Level-1 effects				
Year (Centered)	1.01	0.00	1.01 – 1.02	<0.001
Population (CWC-1)	1.00	0.00	1.00 – 1.00	0.86
Facilities (CWC-1)	0.99	0.00	0.99 – 1.00	0.02
Owner: Private (CWC-1)	1.02	0.03	0.97 – 1.07	0.41
Owner: PPP (CWC-1)	0.99	0.03	0.94 – 1.05	0.73
Servicer: NTNCWS (CWC-1)	1.01	0.02	0.96 – 1.06	0.78
Servicer: TNCWS (CWC-1)	0.93	0.02	0.89 – 0.97	<0.01
Source: Surface (CWC-1)	1.07	0.02	1.03 – 1.11	<0.001
Source: Unknown (CWC-1)	0.98	0.02	0.94 – 1.02	0.23
Level-2 effects				
Population (CWC-2)	1.00	0.00	1.00 – 1.00	0.12
Facilities (CWC-2)	0.98	0.01	0.96 – 1.00	0.03
Owner: Private (CWC-2)	1.60	0.06	1.49 – 1.73	<0.001
Owner: PPP (CWC-2)	1.17	0.25	0.77 – 1.78	0.46
Servicer: NTNCWS (CWC-2)	0.66	0.03	0.60 – 0.73	<0.001
Servicer: TNCWS (CWC-2)	0.43	0.02	0.40 – 0.46	<0.001
Source: Surface (CWC-2)	0.83	0.05	0.75 – 0.93	<0.01
Source: Unknown (CWC-2)	1.06	0.17	0.77 – 1.46	0.74
Level-3 effects				
Population (CGM-3)	0.99	0.01	0.97 – 1.00	0.03
Facilities (CGM-3)	1.13	0.25	0.73 – 1.74	0.59
Owner: Private (CGM-3)	0.66	0.43	0.18 – 2.37	0.52
Owner: PPP (CGM-3)	4.36	10.79	0.03 – 560.22	0.55
Servicer: NTNCWS (CGM-3)	0.96	0.96	0.14 – 6.84	0.97
Servicer: TNCWS (CGM-3)	0.60	0.42	0.15 – 2.38	0.47
Source: Surface (CGM-3)	0.86	0.58	0.23 – 3.20	0.82
Source: Unknown (CGM-3)	0.89	4.01	$0.00 - 6.02e^3$	0.98
Poverty Rate (CGM-3)	1.00	0.00	0.99 – 1.00	<0.001
Pct Republican (CGM-3)	0.98	0.00	0.98 – 0.98	<0.001
Precipitation (CGM-3)	1.00	0.00	1.00 – 1.00	0.38

Note. Total $N = 66,770$.

Estimates, standard errors (*SE*), and 95% confidence intervals (*CI*) are presented as Incidence Rate Ratios.

Table 5

Random Effects for the 4th RIFS Model

	Estimate
Variance components	
Observation-level variance	0.23
System-level variance	1.74
County-level variance	0.29
Variance explained	
Intraclass correlation	0.90
Marginal / Conditional R^2	0.08 / 0.91

Note. Total $N = 66{,}770$.

Figure 13

Estimates of the 4th RIFS Model in Stage 2

Dot-and-Whisker Plot of CWC-1/CWC-2/CGM-3 IRRs
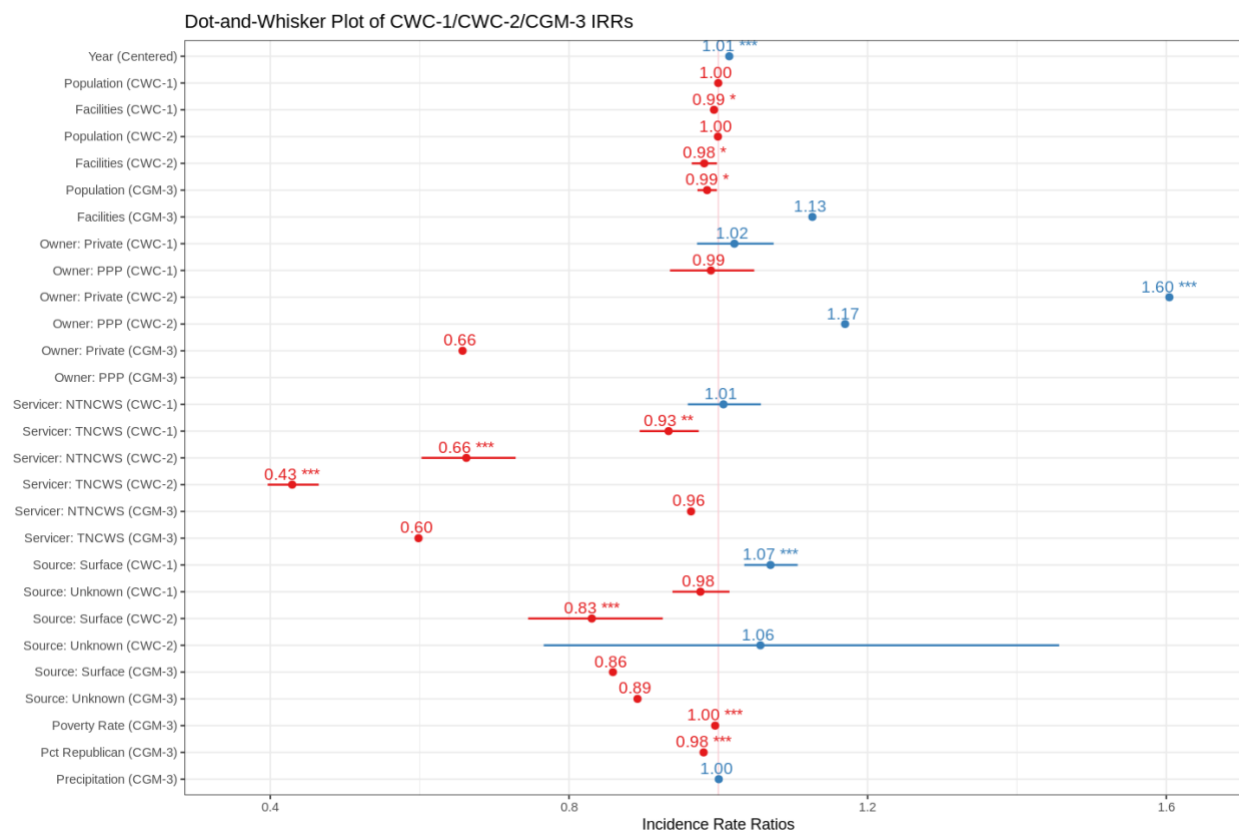
Stage 3: Random Intercept, Random Slope (RIRS)

Models with random intercepts and random slopes were specified in the third stage of progressive model specification. First, a model was fitted where the relationship between the centered time variable and violations was allowed to vary across groups. Group-specific growth or decay trajectories, with respect to regulatory violations, could be captured by this random slope term. However, the model failed to converge properly during maximum likelihood estimation. Convergence failure can occur when a portion of the variance-covariance matrix becomes populated with zeros, a condition known as rank-deficiency. Small or negligible variances are linked to this problem, and lack of convergence here may indicate that there is virtually no variation in the rate of change over time for expected violation counts.

The next model attempted to include all Level-1 and Level-2 centered dummy variables associated with ownership type. Since ownership type is the main variable of interest in this study, the investigation of varying ownership effects across systems was deemed warranted. All Level-1 and Level-2 terms of the categorical variable were included in the random slope specification, in alignment with the guidelines from Rights & Sterba (2023). The second model also faced convergence issues. This estimation difficulty could have been the result of ownership type affecting all systems very similarly, with little to no deviation from the overall mean effect. Put differently, the random slopes and random intercepts may be perfectly or nearly perfectly correlated, introducing multicollinearity and unstable random effects estimation.

Alternative models were fitted that included only Level-1 ownership type terms or only Level-2 ownership type terms. Neither of these models successfully converged toward viable parameter estimations. The specification of a random slope term for other variables could have

been explored, but fitting multiple random slope models incurs a high computational cost. Moreover, the random intercepts already account for over 90% of the variance in expected violation counts and the residual variances of the stage-2 models were minor at 0.23. Therefore, a worthwhile increase in explained variance due to the inclusion of a random slope appeared to be unlikely, as there was little variation left to be explained.

Due to the lack of sufficient within-system variability and the between-system variability already well-accounted for by the random intercept terms, the application of a covariance structure was expected to meet convergence issues as well. This expectation was validated as attempts to include an AR-1 structure or compound symmetry structure led to estimation failures. Additional complexities such as random slopes and covariance structures did not appear to be necessary to model the variation between systems or the correlations within systems.

Stage 4: Same-Level Interactions (SLI)

The fourth stage consisted of fitting three different same-level interaction (SLI) models. The interactions observed were between the Level-2, between-system effect of ownership type and the Level-2, between-system effect of other water system characteristics. The *IRR*s for the fixed effects that were previously interpreted did not change significantly. In fact, the only *IRR* that changed more than 0.01 points was the *IRR* for the Level-2 surface water effect (*IRR* = 0.84, 0.86, and 0.87, respectively), indicating a diminishing effect size. The first SLI model involved an interaction between Level-2 ownership type and Level-2 population served. The *IRR* for this interaction was statistically significant but practically negligible (*IRR* = 1.00; $p = 0.01$). This result suggests that the number of people served by a water system does not modify the effect of ownership type on regulatory compliance.

A second interaction model examined the effect modification potential of water system servicer type. A statistically significant interaction was found between Level-2 private ownership and Level-2 NTNCWS servicer type, suggesting that the between-system effect of ownership type on the expected count of violations differs depending on the water system servicer type. For water systems within a given county that were classified as NTNCWS systems during the entire ten-year period, the impact of exclusively private ownership rather than exclusively public ownership on regulatory compliance was 46% lower compared to systems classified as CWS systems during the entire ten-year period ($IRR = 0.54$; $p < 0.001$).

The potential interaction between ownership type and water source at the second, between-system level was examined in the third model. Interactions between private ownership and surface water source, private ownership and unknown or mixed source, and public-private partnership (PPP) and surface water source were statistically significant ($p = 0.01$, < 0.01, and 0.02, respectively). These interactions suggest that the between-system effect of ownership type on the expected violation counts varied depending on the water source that the system used. Within a given county, when water systems only used surface water during the entire study period, the impact of exclusively private ownership rather than exclusively public ownership on expected violation counts was 32% higher, and the impact of exclusively PPP ownership rather than exclusively public ownership was 647% higher, relative to when water systems only used groundwater ($IRR = 1.32$ and 7.47, respectively). For water systems within a given county that only used unknown or mixed sources, the impact of exclusively private ownership rather than exclusively public ownership on regulatory violations was 175% higher relative to water systems that only used groundwater ($IRR = 2.75$).

The results suggested that surface water usage may amplify the risk of higher violation counts that were found to be linked to greater years spent under private ownership. Surface water usage also appeared to greatly modify the effect of longer PPP ownership. However, the overall mean effect of PPP ownership on regulatory violations was found to be statistically insignificant. With respect to unknown or mixed water sources, greater usage of these sources as a system's primary water supply, may amplify the effect of longer private ownership. However, it is uncommon for a water system to use this type of water source in general and for an unbroken series of years, so the practical importance of this finding may be limited. The random effects, *ICC*, and R^2 measures for all three models remained consistent with those of the last RIFS model in stage 2.

Table 6

Same-Level Interactions (Level-2 Variables)

	Model 1	Model 2	Model 3
Interaction terms			
Owner: Private x Population	0.99*** (0.00)		
Owner: PPP x Population	0.98* (0.01)		
Owner: Private x Servicer: NTNCWS		0.54*** (0.06)	
Owner: PPP x Servicer: NTNCWS		2.79 (1.66)	
Owner: Private x Servicer: TNCWS		0.84 (0.07)	
Owner: PPP x Servicer: TNCWS		0.81 (0.46)	
Owner: Private x Source: Surface			1.32* (0.14)
Owner: PPP x Source: Surface			7.47* (6.56)
Owner: Private x Source: Unknown			2.75** (1.03)
Owner: PPP x Source: Unknown			26.47 (65.11)
Variance components			
Observation-level variance	0.23	0.23	0.23
System-level variance	1.74	1.73	1.74
County-level variance	0.29	0.29	0.29
Variance explained			
Intraclass correlation	0.90	0.90	0.90
Marginal / Conditional R^2	0.08 / 0.91	0.08 / 0.91	0.08 / 0.91

Note. Total $N = 66,770$. ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Estimates and standard errors (in parentheses) are presented as Incidence Rate Ratios.

Figure 14

Comparison of Stage-4 Model Estimates

Dot-and-Whisker Plots of SLI Models
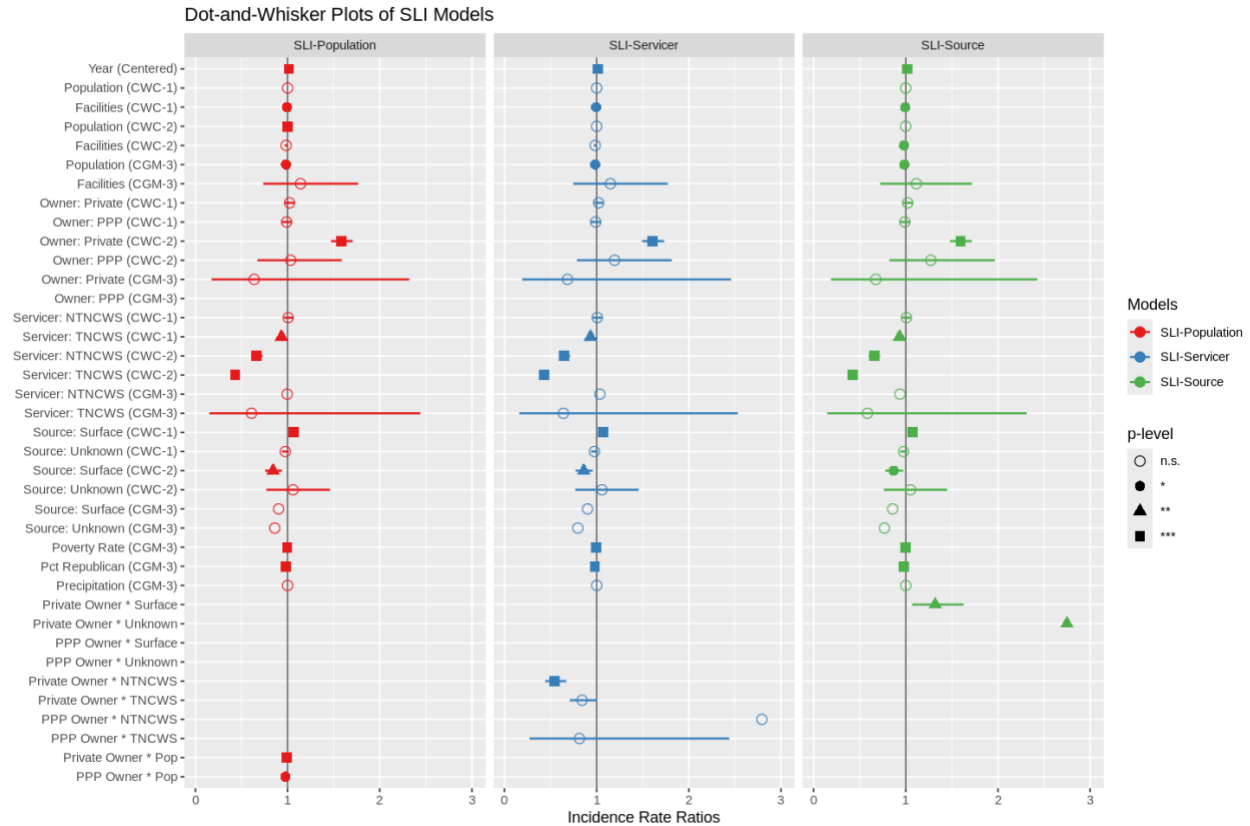
Stage 5: Cross-Level Interactions (CLI)

Two cross-level interactions (CLI) were investigated in the fifth stage of the analysis. First, a potential interaction between Level-2, centered-within-cluster ownership type and Level-3, grand-mean-centered poverty rate was explored. Statistically significant but practically negligible estimates were observed (*IRR* = 1.01 and 1.05, respectively and *p* < 0.001 for both), suggesting that the impact of longer private ownership or longer PPP ownership on violation counts do not vary based on counties' poverty rate deviations from the state mean. For a one-percent increase in the county-level poverty rate above the state mean, the impact of exclusively private ownership versus exclusively public ownership on regulatory compliance increases by

1%, and the impact of exclusively PPP ownership increases by 5%. Therefore, a county-level poverty rate that exceeds the state mean does not appear to substantially modify the compliance risks associated with more years spent under private ownership.

A similar pattern of little-to-no modification was observed in the model that included the interaction between Level-2 ownership type and Level-3 percentage of county registered voters who registered Republican. For a two-percent increase in the county-level percentage above the state mean, the impact of exclusively private ownership versus exclusively public ownership on regulatory compliance increases by 1%, and the impact of exclusively PPP ownership increases by 9% ($IRR = 1.01$ and 1.09, respectively; $p < 0.01$ and < 0.001). In other words, having a percentage of Republican voters above the state mean does not appear to substantially modify the effects of longer private ownership. As Level-2 PPP ownership had no statistically significant effect on the expected count of violations on its own, the 9% effect modification may not have much practical significance. The random effects, ICC, and R^2 measures for the two CLI models remained consistent with the random effects, ICC, and R^2 measures of previous three-level models.

Table 7

Cross-Level Interactions (Level-2 and Level-3 Variables)

	Model 1	Model 2
Interaction terms		
Owner: Private x Poverty Rate	1.01*** (0.00)	
Owner: PPP x Poverty Rate	1.05*** (0.01)	
Owner: Private x Pct Republican		1.01** (0.00)
Owner: PPP x Pct Republican		1.09*** (0.02)
Variance components		
Observation-level variance	0.23	0.23
System-level variance	1.75	1.75
County-level variance	0.29	0.29
Variance explained		
Intraclass correlation	0.90	0.90
Marginal / Conditional R^2	0.08 / 0.91	0.08 / 0.91

Note. Total $N = 66,770$. ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

Estimates and standard errors (in parentheses) are presented as Incidence Rate Ratios.

Figure 15

Comparison of Stage-5 Model Estimates
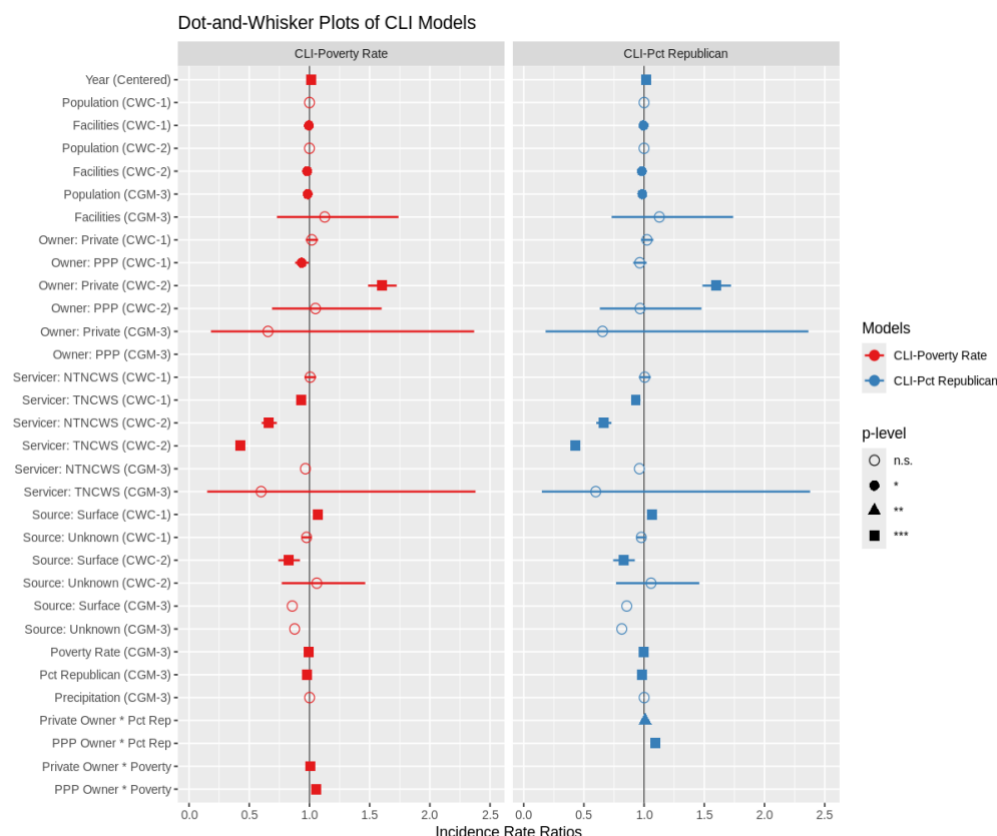
Residual Diagnostics

Residual diagnostics using scaled quantile residuals (SQR) were performed with the highest-complexity RIFS model from stage 2. This model contained the system-specific and county-specific random intercepts, all fixed effects for levels 1, 2, and 3, but no random slopes, covariance structures, or interactions between fixed effects. The choice of model for the residual analysis was made in the interest of utilizing the best representation of the data, identified during progressive model specification, while also avoiding unnecessary complexity. The quantile-quantile (QQ) plot in Figure 14 compares the distribution of the scaled quantile residuals to a theoretical uniform distribution. Deviations from the diagonal reference line indicate potential

departures from the distribution expected under correct model specification. In this case, an unexpectedly large number of residuals are concentrated above the diagonal line in the lower-to-mid range. Additionally, a slight deviation below the diagonal line at the right end of the plot indicates that there are less residuals than expected at the very high end of the distribution. These observations suggest a general pattern of non-uniformity over a wide range of the scaled quantile residuals. The overall degree of deviation from the expected uniform distribution is indicative of model misspecification, which may negatively impact subsequent estimation and inference.

Figure 16

*QQ Plot of Scaled Quantile Residuals for Testing Uniformity*
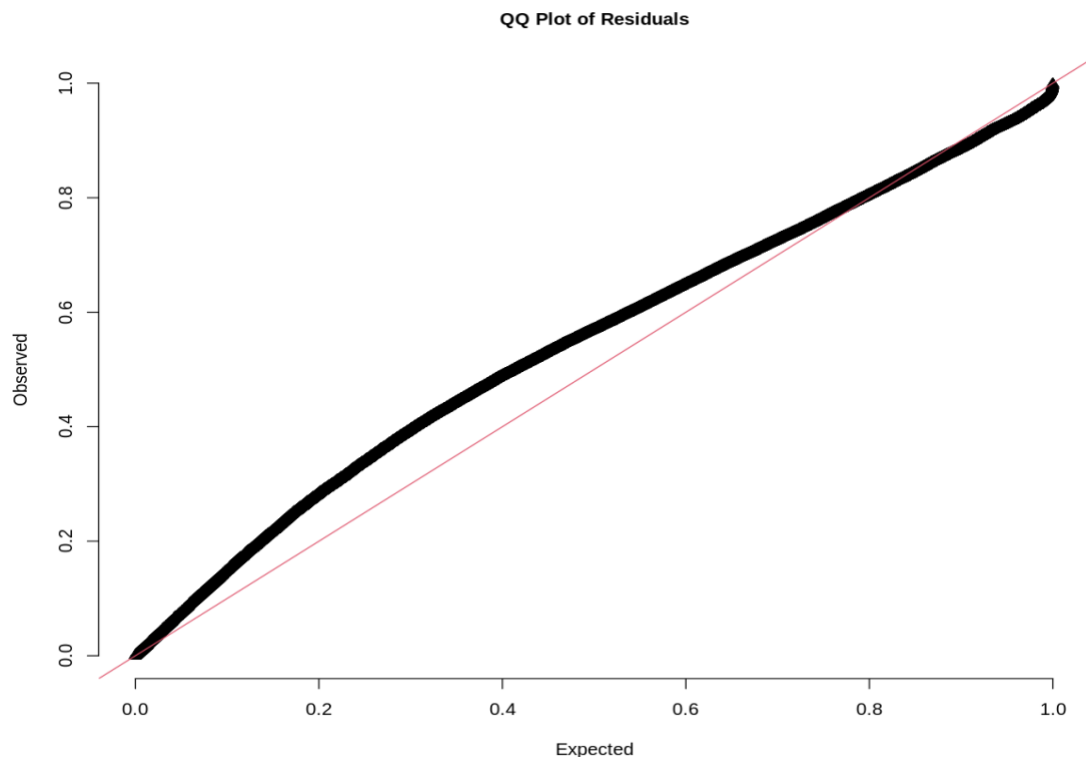
In Figure 15, the plot of scaled quantile residuals versus rank-transformed predicted values is shown. Correct model specification is indicated when the plotted quantile regression

lines fit closely to the flat reference lines representative of the expected uniform distribution. However, the plotted quantile regression lines show consistent deviation from their respective reference lines, providing evidence of non-uniformity. Moreover, a distinct, curvilinear pattern can be observed for all the quantiles, indicating the presence of systematic non-linearity that has not been accounted for by the model. In the leftmost area of the figure, the plotted lines deviate above their reference lines more noticeably, meaning that there is a higher frequency of residuals than expected for lower predicted values. At the right end of the plot, the quantile regression lines move closer to expected behavior, but then curve upward and reach varying heights depending on the quantile. For higher predicted values, a higher-than-expected frequency of residuals is most prevalent at the 0.25 quantile, becomes less pronounced at the 0.50 quantile, and then reverses to a lower-than-expected frequency at the 0.75 quantile. The non-uniform, non-linear behavior observed in the diagnostic plot indicates that the model may be misspecified. There may be missing variables, non-linear relationships, or structural components (e.g. an integrated zero-inflation or hurdle model) that, if identified and included, could improve model fit and validity.

Figure 17

Plot of Scaled Quantile Residuals vs. Rank-Transformed Predictions
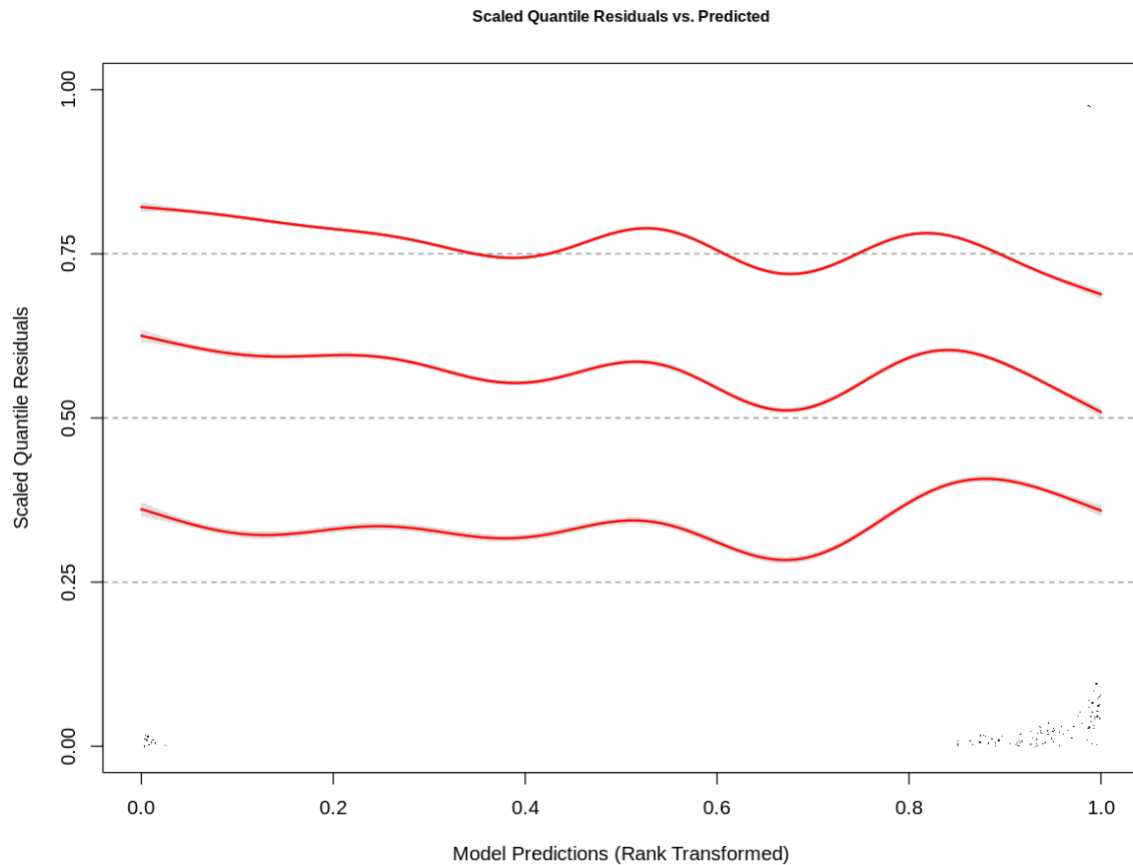
Predictive Modeling Results

A predictive modeling analysis was conducted to observe the relative importance of ownership type for predicting the number of regulatory violations committed by water systems. This task was carried out with a standard cross-sectional approach rather than with a time series approach. This decision was made in lieu of the limited number of observations per water system, which was presumed to be insufficient for most learning algorithms to adequately model. A hybrid model, mixing elements of a multilevel negative binomial mixed effects model and a

gradient boosting tree model, was created with the GPBoost package in Python and employed for this analysis. The model's performance on unseen data and different types of feature importance were examined to address the research question.

Model Training and Validation

The variables used in the prior explanatory analysis were also used as features in this machine learning task, in addition to appropriately centered versions of the wholesaler status and county-level average temperature variables. The mixed effects gradient boosting tree model (ME-GBT) was defined with a negative binomial probability distribution, system-specific and county-specific random intercepts, and the Nelder-Mead optimizer for maximum likelihood estimation. The choice of optimizer was recommended by the creator of the GPBoost package to reduce the computational burden of training GPBoost models.

The longitudinal panel dataset was partitioned into training, validation, and testing sets with an 80/10/10 split. Special care was given to ensuring that observations from one entity remained within one set, rather than being distributed among the different sets. The validation set was used to impose an early stopping condition on the training process, in the interest of mitigating the computational cost of training. After 50 boosting rounds without improvement in negative log-likelihood when tested on the validation set, training was set to terminate, which is precisely what occurred during the training process. The estimated random effect variances for water systems and counties were 1.86 and 0.48, respectively. These parameters are comparable to the random effect variances estimated by the explanatory models.

Prediction and Importance Evaluation

The ME-GBT model obtained an *MSE* of 200.25 and an *RMSE* of 14.15 when evaluated on the testing set. However, since no other learning algorithms were used for this task, it cannot be said whether this model underperforms or overperforms relative to other types of models. The plots of split-based feature importances and gain-based feature importances were examined after evaluating on the test set. Only features that obtained a non-zero importance score were included in the visualizations. The plot of split-based feature importances shows that the centered year variable was the feature that was used most frequently to split the data ($fi = 6$). If the model identified the observation year as the most important feature for predicting violation counts, then there may be unobserved factors associated with certain years that could influence regulatory compliance. On the other hand, this result may have occurred because observations within water systems are highly correlated with one another and generally display low variability across the study period. Additionally, missing or insufficiently informative variables may have led to these results. Level-2 population served, Level-2 TNCWS classification, and Level-3 poverty rate were the next-most important features in terms of split frequency ($fi = 3$). The feature representing Level-3 PPP ownership had a feature importance score of 2, but no other ownership type features achieved a non-zero feature importance score.

Figure 18

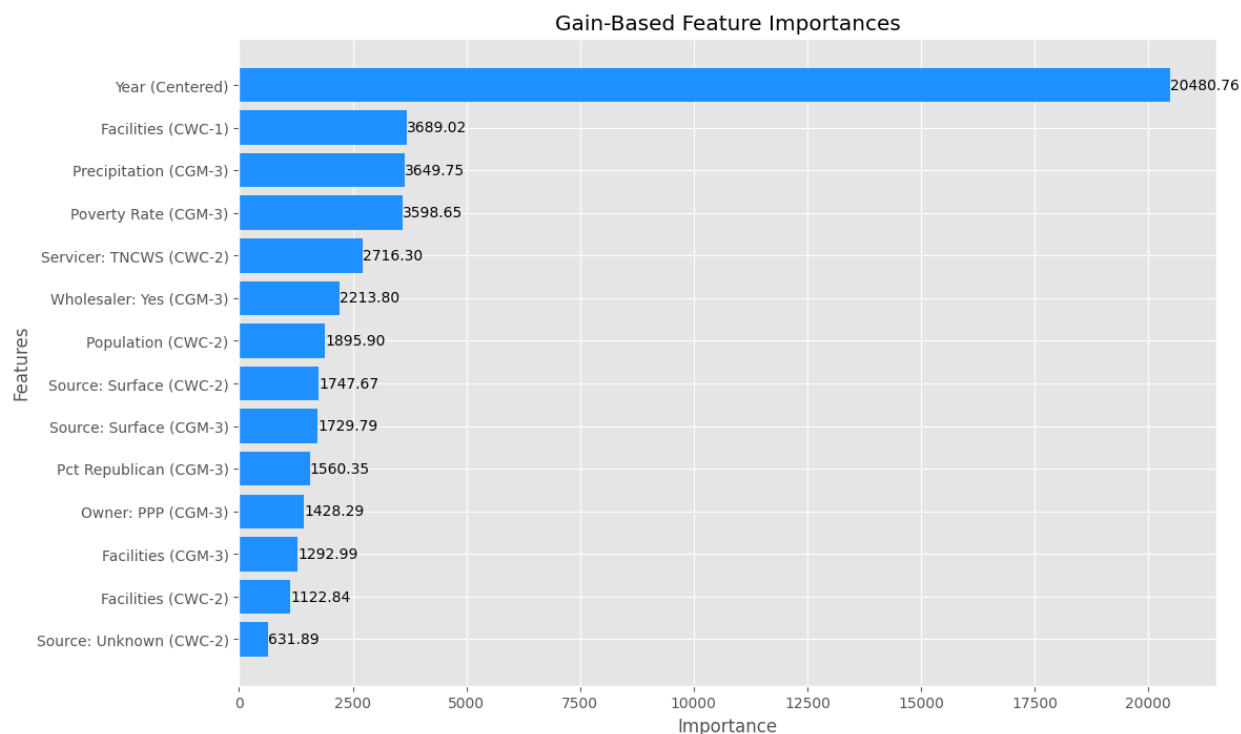*Plot of Split-Based Feature Importance Scores*
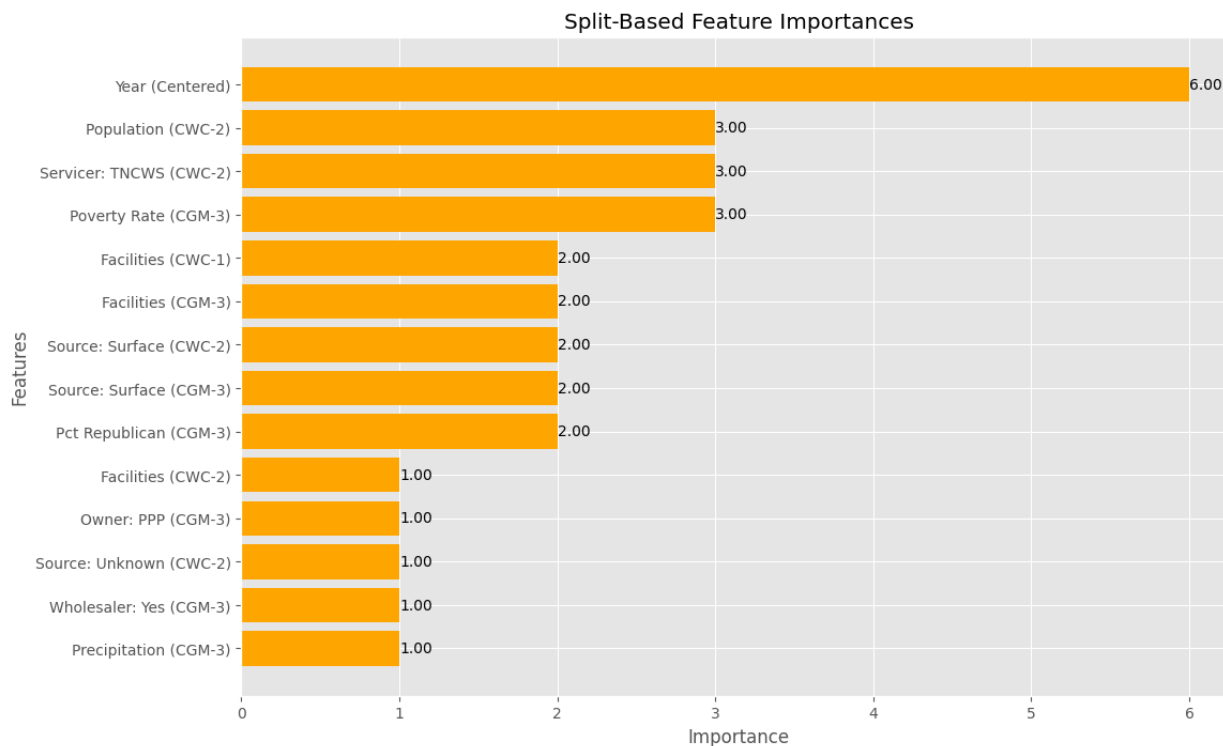
The plot of gain-based feature importances showed a similar pattern in terms of the features that were identified as important. The year was again found to be the most important feature. In the context of gain-based feature importance, this means that splitting based on the observation year led to the most model improvement, on average (fi = 20,480.76). During the explanatory analysis, adding a random slope for year resulted in convergence issues, suggesting the possibility of a near-zero variance parameter for the random slope and implying that the rate of change in violations over time did not vary considerably. Moreover, the effect size for the centered year variable was extremely small, with an estimated 1 to 2% increase in violations from one year to the next, on average. These findings support the notion that violation counts

may remain relatively stable over time within the majority of water systems. Therefore, it is entirely plausible that the year variable would be the most relevant feature in a predictive context. If violation counts remain stable over time, and time is included as a feature, then the model may rely heavily on this feature to minimize the error in its predictions. The three most important features after the time variable were the number of facilities at the observation level, the average precipitation at the county level, and the poverty rate at the county level. Poverty rate was also identified as an influential feature via split-based feature importance. In addition, many of the same features that achieved non-zero split-based feature importance also achieved non-zero gain-based feature importance, suggesting a degree of consensus among the different feature importance methods. However, ownership type, with the exception of county-level PPP ownership, did not contribute in any meaningful capacity to the model predictions.

Figure 19

Plot of Gain-Based Feature Importance Scores

Gain-Based Feature Importances

Feature	Importance
Year (Centered)	20480.76
Facilities (CWC-1)	3689.02
Precipitation (CGM-3)	3649.75
Poverty Rate (CGM-3)	3598.65
Servicer: TNCWS (CWC-2)	2716.30
Wholesaler: Yes (CGM-3)	2213.80
Population (CWC-2)	1895.90
Source: Surface (CWC-2)	1747.67
Source: Surface (CGM-3)	1729.79
Pct Republican (CGM-3)	1560.35
Owner: PPP (CGM-3)	1428.29
Facilities (CGM-3)	1292.99
Facilities (CWC-2)	1122.84
Source: Unknown (CWC-2)	631.89

Summary

This chapter reported the results of two analytical endeavors. First, an explanatory analysis was performed to assess whether a relationship between water system ownership type and regulatory violations was present. Other potential risk factors and interactions were analyzed as well. This explanatory analysis was conducted using a multilevel negative binomial mixed effects model to account for group-level heterogeneity and non-independence, in addition to the potential overdispersion surrounding the count-based dependent variable. Variables were decomposed via centering techniques to isolate their level-specific effects. A five-stage process for progressive model specification was implemented to obtain a more in-depth analysis of the

multilevel hierarchical structure and a reliable assessment of estimation stability across varying specifications.

The results of the analysis showed that a greater number of years spent under private ownership rather than public ownership is associated with an increase in water system regulatory violations. In particular, water systems within a given county that spent all ten years under private ownership rather than public ownership were expected to have 60% more violations, on average. Additionally, at the between-system level, non-CWS classification was associated with lower expected violation counts, while surface water usage was associated with higher expected violation counts. Contrary to expectations, when poverty rates and the percentage of the Republican voter base in a county increased relative to the state mean, no substantial differences in regulatory violation rates were observed. Interactions were investigated, finding that servicer type and water source acted as modifiers of the between-system private ownership effect.

The random effect parameters revealed that there was substantial variability at the system level compared to the county level. Furthermore, a full 90 to 91% of explained variance was accounted for by the mere inclusion of random intercept terms, with remaining residual variance being reduced to 0.23. The inclusion of fixed effects added virtually no explanatory power to the model, suggesting that important variables may be unaccounted for. Examination of scaled quantile residual diagnostics indicated potential misspecification as the model failed to align with the expected uniform distribution of the theoretical quantiles. Overall, the residual diagnostics showed that there was room for improvement in terms of model specification, and lent further credibility to the notion that more-relevant variables may be necessary to enhance model fit and validity.

After the explanatory analysis, a predictive analysis was conducted to investigate the relative importance of ownership type within a prediction-based context. A mixed effects gradient boosting tree model was trained to predict violation counts for unseen water systems within the same timeframe. An *MSE* of 200.25 and *RMSE* of 14.15 were obtained when evaluating the model on the holdout test set. Split-based feature importances and gain-based feature importances appeared to favor the centered time variable as the most useful feature for splitting and predictions. The role of time and its relationship to violation counts within systems was considered in more detail. In contrast, ownership type did not play a meaningful role in model predictions, as shown in the results of both feature importance methods.

Chapter 5: Conclusions

Changing ownership of public water systems in California has sparked debate about the potential advantages and disadvantages of privatizing water utilities, particularly concerning the ability to maintain regulatory standards and provide safe drinking water to residents. This study aimed to investigate the impact of privatization on water quality compliance in California's public water systems, while assessing demographic and geographic factors that may affect compliance outside of ownership type alone. By employing negative binomial mixed effects models and a multilevel modeling framework, the study analyzed data from the Environmental Protection Agency, the U.S. Census Bureau, the National Oceanic and Atmospheric Administration, and the Office of the California Secretary of State over a ten year period.

The results demonstrated that private ownership of water systems is associated with a higher incidence of regulatory violations compared to public ownership. Specifically, within a given county, water systems that were privately owned for all ten years of the study had 60% more violations on average than water systems that were publicly owned for all ten years. Additionally, factors such as water system servicer type and primary water source significantly influenced expected violation rates and the effects of ownership type. The random effect parameters revealed substantial variability at the system level, suggesting that individual water system characteristics play a crucial role in determining compliance outcomes. Furthermore, the inclusion of fixed effects added virtually no explanatory power to the model, indicating potential unaccounted variables that could further improve model fit and validity. The predictive analysis, conducted using a mixed effects gradient boosting tree model, showed that ownership type was not a meaningful predictor, with time-related variables taking precedence in terms of feature importance.

Limitations

While this study provides valuable insights into water contamination and water system ownership, several limitations must be acknowledged to contextualize the findings and guide future research. First, the lack of cross-validation in the machine learning portion significantly limits the robustness of the predictive model. Without cross-validation, the model's performance on unseen data remains uncertain. The model's generalizability cannot be confidently assessed, which may result in overfitting where the model performs well on the training data but poorly on new, independent datasets. Consequently, the conclusions drawn from the model's predictions are less reliable, potentially affecting the study's overall validity and applicability to broader contexts.

Next, it is possible that using total violations as a proxy for system compliance obscures differences between private and public water systems with respect to specific contaminants or broader classes of contaminants. Currently, the EPA sets legal limits on over 90 different contaminants (Environmental Protection Agency, 2023) and while including each would result in an unwieldy dataset (at best) it is possible to either place these contaminants into broader categories or only choose the most prevalent. The EPA does so itself at times, delineating contaminants into either chemical or microbial contaminants, or into six separate categories (microorganisms, disinfectants, etc.) depending on the context. In addition, contaminant data could be analyzed with regards to acute vs. chronic exposure risks. For example, short-term exposure to *Legionella* and long-term exposure to elevated arsenic levels are both serious concerns for consumers. Yet the potential effects of private vs. public ownership status on these contaminant levels remain unexplored when using the single variable of total violations. Specific contaminants of concern are not examined in this dataset, most notably nitrate. The introduction

of synthetic fertilizers in the 1940s has led to continually higher concentrations of nitrate levels in groundwater (Burow et al., 2013). California has long been the most productive agricultural state in the country, and several regions have shown increases in nitrate concentrations consistent with the increased use of synthetic fertilizers over time.

Thirdly, recent work by Beecher et al. (2020) has highlighted troublesome aspects of the Safe Drinking Water Information System typing and coding process. The authors note several issues that show water system data is often more nuanced than typically presented. For instance, about 19% of community water systems do not actually produce their own water, instead purchasing treated water from other systems. Variability in state oversight and reporting consistency raise serious questions regarding data quality. Lastly, the research notes that the "private" category often includes not-for-profit systems, and the public/private category often acts as a kind of "catchall" category for systems where ownership is unknown or not coded at all.

Finally, some factors that make California a uniquely interesting topic of research may limit the broader applicability of this specific model. The use of yearly precipitation totals as a variable is the most notable example. Precipitation variability in California is extremely high compared to the rest of the United States, even the Pacific Northwest states of Oregon and Washington (California Department of Water Resources, 2023). Yearly precipitation variability in many areas of California has a coefficient of variation as high as 0.6 or 0.7. By point of comparison, these values east of the Mississippi rarely exceed 0.2. However, it is entirely possible other parts of the world experience this kind of variability and would find this data useful.

Implications for Stakeholders

The work adds to the relatively sparse body of research on differences in water contamination levels between publicly and privately owned water systems. Policymakers in California have a long history of implementing water data and research into legislation. State agencies, particularly the California Department of Water Resources, have an enormous body of publicly available research. Since 1957 the state has developed and published a state water plan every five years. Governmental agencies and politicians have a long track record of using and incorporating water quality data.

Water system research is potentially valuable not only in California, but in developing nations as well. In the United States, relatively few community water systems report violations (about 7-8% per year), but this level of overall quality is only possible because of many decades of research and experimentation. This knowledge base currently informs much of the work being done in developing countries, and as researchers continue to produce high quality research here in America, these findings can potentially be applied overseas as well. Tragically, the two main waterborne diseases of 19th century America, cholera and typhoid, continue to affect millions around the world. Poor quality data makes an exact measurement of disease incidence impossible, but broad figures are available. A recent estimate (Ganesan et al., 2020) of yearly cholera incidence showed 1.3-4.0 million cases and 21,000-143,000 deaths. In 2015, cases were reported in 42 countries, and cholera occurs endemically in many parts of South Asia and sub-Saharan Africa. Typhoid fever is currently an even greater threat to people's health worldwide. A 2020 estimate (Khanam et al., 2020) of the global burden of typhoid fever had a total of 9 million cases and 110,000 deaths per year. The researchers note this is an undercount, as only confirmed cases from Africa and Asia were included in the overall total. Several times the present research

has highlighted several ways that California was able to elude water contamination issues faced by earlier developing countries and cities. While several areas of the developing world have not been as fortunate with regards to many waterborne diseases, there is no reason why the hard-won lessons and knowledge that benefit America cannot be used to alleviate the devastating effects of water contamination in other countries.

Future Research

There are opportunities for further research on water quality and water system ownership status. In the near future, it is highly likely that renewed efforts to replace pipes containing lead will emerge in light of the recent Flint water crisis. Elevated levels of lead are not caused by poor sanitation or filtration policies, but instead by water being transported through these pipes. The problem can therefore only be addressed by replacing existing infrastructure, a challenge that public and private companies may address (or not) in different ways. Researchers would do well to pay close attention to how these systems handle this problem, especially considering lead contamination has recently captured public attention.

As mentioned previously, there is limited understanding on how public versus private water utilities are preparing for, and adapting to, the impacts of climate change. Climate change mindfulness is especially important for California, as reduced yearly snowfall and an increased number of wildfires have already significantly altered water availability and delivery. While future effects resulting from climate change are yet to be determined, there is the potential for devastating changes to the state's water situation.

Conclusion

In conclusion, this study provides significant insights into the impact of privatization on water quality compliance in California. The research confirmed that private water systems are more likely to incur regulatory violations over time, potentially affirming concerns about the efficacy of privatized water management. However, the study also highlights the complexity of predicting water quality compliance, indicating that other variables beyond ownership type must be considered. The implications of these findings are substantial for policymakers and stakeholders in the water management sector. Understanding that privatization correlates with higher violation rates underscores the need for stringent regulatory oversight and possible reconsideration of privatization policies. Moreover, the limited predictive power of ownership type suggests that multifaceted strategies, considering various demographic and geographic factors, are essential for improving water quality and regulatory compliance.

This study contributes to the broader discourse on water management by providing empirical evidence on the consequences of privatization. Future research should explore additional variables that influence water quality compliance and further refine predictive models to enhance their accuracy and reliability. Through these efforts, policymakers can be better equipped to make informed decisions that prioritize public health and safety in the management of water resources.

References

Acquah, S., & Allaire, M. (2023). Disparities in drinking water quality: Evidence from California. *Water Policy, 25*(2), 69-86. https://doi.org/10.2166/wp.2023.068

Adler, A. I., & Painsky, A. (2022). Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy, 24*(5), 687. https://doi.org/10.3390/e24050687

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management, 39*(6), 1490-1528. https://doi.org/10.1177/0149206313478188

Allaire, M., Wu, H., & Lall, U. (2018). National trends in drinking water quality violations. *Proceedings of the National Academy of Sciences*, *115*(9), 2078–2083. https://doi.org/10.1073/pnas.1719805115

Anthony, C. (2009). *Water Privatization Trends in the United States: Human Rights, National Security, and Public Stewardship*. *33*(3), 785.

Arif, S., & MacNeil, M. A. (2022). Model selection isn't causal inference. *Authorea Preprints*. https://doi.org/10.22541/au.164440121.18799387/v1

Ashley, P. (1906). The water, gas, and electric light supply of London. The ANNALS of the American Academy of Political and Social Science, 27(1), 20-36.

ATSDR. (2007b). Toxicological Profile for Arsenic. Agency for Toxic Substances and Disease Registry, U.S. Department of Health and Human Services, Atlanta, GA. Available at: http://www.atsdr.cdc.gov/ToxProfiles/tp2.pdf.

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology, 37*(1), 13-20. https://doi.org/10.1017/S0048577200980648

Baker, M. N. (1892). The manual of American water-works.

Baker, M. N. (1915). Municipal ownership and operation of water works. The Annals of the American Academy of Political and Social Science, 57(1), 279-281.

Barnes, D. S. (2006). The great stink of Paris and the nineteenth-century struggle against filth and germs. JHU Press.

Bates, S., Hastie, T., & Tibshirani, R. (2023). Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association, 119*(546), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

Beach, B. (2022). Water infrastructure and health in US cities. *Regional Science and Urban Economics*, *94*, 103674.

Beecher, J., Redican, K., & Kolioupoulos, M. (2020). (Mis) Classification of water systems in the United States. *Available at SSRN 3627915*.

Bel, G. (2020). Public versus private water delivery, remunicipalization and water tariffs. *Utilities Policy*, *62*, 100982. https://doi.org/10.1016/j.jup.2019.100982

Bel, G., & Warner, M. (2008). Does privatization of solid waste and water services reduce costs? A review of empirical studies. *Resources, Conservation and Recycling*, *52*(12), 1337–1348. https://doi.org/10.1016/j.resconrec.2008.07.014

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*, 1937-1967. https://doi.org/10.1007/s10462-020-09896-5

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127-135. doi: 10.1016/j.tree.2008.10.008. PMID: 19185386.

Bouwer, H. (2000). Integrated water management: Emerging issues and challenges. *Agricultural Water Management, 45*(3), 217-228. https://doi.org/10.1016/S0378-3774(00)00092-5

Brincks, A. M., Enders, C. K., Llabre, M. M., Bulotsky-Shearer, R. J., Prado, G., & Feaster, D. J. (2017). Centering Predictor Variables in Three-Level Contextual Models. *Multivariate Behavioral Research, 52*(2), 149–163. https://doi.org/10.1080/00273171.2016.1256753

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378-400. https://doi.org/10.32614/RJ-2017-066

Brooks, R. C. (2004). Privatization of government services: An overview and review of the literature. *Journal of Public Budgeting, Accounting & Financial Management, 16*(4), 467-491. https://doi.org/10.1108/JPBAFM-16-04-2004-B001

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920960351

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104*(3), 396. https://doi.org/10.1037/0033-2909.104.3.396

Budds, J., & McGranahan, G. (2003). Are the debates on water privatization missing the point? Experiences from Africa, Asia and Latin America. *Environment and urbanization*, *15*(2), 87-114.

Burow, K. R., Jurgens, B. C., Belitz, K., & Dubrovsky, N. M. (2013). Assessment of regional change in nitrate concentrations in groundwater in the Central Valley, California, USA, 1950s–2000s. *Environmental earth sciences*, *69*(8), 2609-2621.

California Department of Food and Agriculture. (2023). California Agricultural Statistics Review 2022-2023. https://www.cdfa.ca.gov/Statistics/PDFs/2022-2023_california_agricultural_statistics_review.pdf

California Department of Water Resources. (1957). The California Water Plan. https://archive.org/details/californiawaterp03cali/page/n3/mode/2up

California Department of Water Resources. (2021). California's groundwater: Update 2020. https://data.cnra.ca.gov/dataset/calgw_update2020

California Department of Water Resources. (2023). California water plan update 2023. https://water.ca.gov/Programs/California-Water-Plan/Update-2023

California Office of Environmental Health Hazard Assessment. (2022, November 1).

Precipitation. OEHHA. https://oehha.ca.gov/climate-change/epic-2022/changes-

climate/precipitation

California Secretary of State. (n.d.-a). About Us. California Secretary of State. Retrieved June 1,

2024, from https://www.sos.ca.gov/administration

California Secretary of State. (n.d.-b). Voter Registration Statistics. California Secretary of State.

Retrieved June 1, 2024, from https://sos.ca.gov/elections/voter-registration/voter-registration-

statistics

Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data* (Vol. 53).

Cambridge University Press. https://doi.org/10.1017/CBO9781139013567

Campbell, W. A. B. (1952). Methaemoglobinaemia due to nitrates in well-water. British medical

journal, 2(4780), 371.

Cerqueira, V., Torgo, L., & Soares, C. (2019). Machine learning vs statistical methods for time

series forecasting: Size matters. *arXiv Preprint*, arXiv:1909.13316.

https://doi.org/10.48550/arXiv.1909.13316

Chen, C. J., Chuang, Y. C., Lin, T. M., & Wu, H. Y. (1985). Malignant neoplasms among

residents of a blackfoot disease-endemic area in Taiwan: high-arsenic artesian well water and

cancers. Cancer research, 45(11_Part_2), 5895-5899.

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine, 16*(20), 2349-2380. doi: 10.1002/(sici)1097-0258(19971030)16:20<2349::aid-sim667>3.0.co;2-e. PMID: 9351170.

Comly, H. H. (1945). Cyanosis in infants caused by nitrates in well water. Journal of the American Medical Association, 129(2), 112-116.

Cook, S. F. (1955). The epidemic of 1830-1833 in California and Oregon (Vol. 43, No. 1). University of California Press.

Cook, S. F. (1976). The population of the California Indians, 1769-1970. (No Title).

Correia, S. P. D. (2023). *Count mixed-effects regression models in parasite ecology*. Universidade do Porto. https://hdl.handle.net/10216/155274

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press.

https://doi.org/10.1017/CBO9780511813559

Crawford, R. (1906). Glasgow's Experience with Municipal Ownership and Operation: Water, Gas, Electricity and Street Railways. The Annals of the American Academy of Political and Social Science, 27(1), 1-19.

Creswell, J. W., & Creswell, J. D. (2017). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). Sage.

Daly, H. E. (1993). Free market environmentalism. *Ecological Economics*, *7*(2), 173–176. https://doi.org/10.1016/0921-8009(93)90052-8

De Roos, A. J., Kondo, M. C., Robinson, L. F., Rai, A., Ryan, M., Haas, C. N., Lojo, J., & Fagliano, J. A. (2020). Heavy precipitation, drinking water source, and acute gastrointestinal illness in Philadelphia, 2015-2017. *PloS ONE, 15*(2), e0229258. https://doi.org/10.1371/journal.pone.0229258

Denhard, A., Bandyopadhyay, S., Habte, A., & Sengupta, M. (2021). Evaluation of time-series gap-filling methods for solar irradiance applications (No. NREL/TP-5D00-79987). National Renewable Energy Lab. (NREL). https://doi.org/10.2172/1826664

Diggle, P. (2002). *Analysis of Longitudinal Data*. Oxford University Press. https://doi.org/10.1002/sim.1701

Dobson, A. J. (2013). *Introduction to Statistical Modelling*. Springer.

Dobson, A. J., & Barnett, A. G. (2018). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC.

Douglas, T. J. (1976). Safe Drinking Water Act of 1974--History and Critique. Environmental Affairs, 5, 501.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics, 5*(3), 236-244. https://doi.org/10.2307/1390802

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Eutsler, R. B. (1939). Public and private ownership of water supply utilities. The ANNALS of the American Academy of Political and Social Science, 201(1), 89-95.

Fan, A. M., & Steinberg, V. E. (1996). Health implications of nitrate and nitrite in drinking water: an update on methemoglobinemia occurrence and reproductive and developmental toxicity. Regulatory toxicology and pharmacology, 23(1), 35-43.

Fawns, H. T., & Aldridge, A. G. V. (1954). Methaemoglobinaemia due to nitrates and nitrites in drinking-water. British Medical Journal, 2(4887), 575.

Fothergill, A. (1812). On the Poison of Lead. The Belfast Monthly Magazine, 8(44), 168-176.

Fu, G., Liu, P., & Swallow, S. K. (2020). Effectiveness of Public versus Private Ownership: Violations of the Safe Drinking Water Act (SDWA). *Agricultural and Resource Economics Review*, *49*(2), 291–320. https://doi.org/10.1017/age.2020.4

Funatogawa, I., & Funatogawa, T. (2018). *Longitudinal Data Analysis: Autoregressive Linear Mixed Effects Models*. Springer Singapore. https://doi.org/10.1007/978-981-10-0077-5

Gałecki, A., & Burzykowski, T. (2013). Linear mixed-effects model. In *Linear Mixed-Effects Models Using R* (pp. 245-273). Springer New York. https://doi.org/10.1007/978-1-4614-3900-4

Ganesan, D., Gupta, S. S., & Legros, D. (2020). Cholera surveillance and estimation of burden of cholera. *Vaccine*, *38*, A13-A17.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Goluboff, N. (1948). Methaemoglobinaemia in an infant. Canadian Medical Association Journal, 58(6), 601.

Greiner, P. T. (2016). Social Drivers of Water Utility Privatization in the United States: An Examination of the Presence of Variegated Neoliberal Strategies in the Water Utility Sector. *Rural Sociology*, *81*(3), 387–406. https://doi.org/10.1111/ruso.12099

Groenwold, R. H., & Dekkers, O. M. (2023). Is it a risk factor, a predictor, or even both? The multiple faces of multivariable regression analysis. *European Journal of Endocrinology, 188*(1), E1-E4. https://doi.org/10.1093/ejendo/lvac012

Harris, R. H., & Brecher, E. M. (1974). Is the water safe to drink? Part I: The problem. Part II: How to make it safe. Part III: What you can do. Consumer Reports, 436(538), 623.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press. https://doi.org/10.1017/CBO9780511973420

Hillier, J., & Bell, S. (2010). The 'genius of place': mitigating stench in the New Palace of Westminster before the Great Stink. The London Journal, 35(1), 22-38.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*(5), 623–641. https://doi.org/10.1016/S0149-2063(99)80077-4

Hughes, M. F., Beck, B. D., Chen, Y., Lewis, A. S., & Thomas, D. J. (2011). Arsenic exposure and toxicology: a historical perspective. Toxicological sciences, 123(2), 305-332.

Hyatt, E. (1925). Control of Appropriations of Water in California. Journal (American Water Works Association), 13(2), 125-144.

Innes, R., & Mitra, A. (2015). Parties, politics, and regulation: Evidence from clean air act enforcement. *Economic Inquiry, 53*(1), 522-539. https://doi.org/10.1111/ecin.12142

Izaguirre, A. K. (2003). *Private participation in infrastructure: trends in developing countries in 1990-2001: energy, telecommunications, transport, water. World Bank Publications.*

Jonasson, M. E., & Afshari, R. (2018). Historical documentation of lead toxicity prior to the 20th century in English literature. Human & Experimental Toxicology, 37(8), 775-788.

Kazemitabar, J., Amini, A., Bloniarz, A., & Talwalkar, A. S. (2017). Variable importance using decision trees. In *Advances in Neural Information Processing Systems* (Vol. 30).

Kempe, M. (2006). New England Water Supplies–A Brief History: 385 Years of Drinking Water, 125 Years of New England Water Works Association. *Journal of the New England Water Works Association*, *120*(3).

Khanam, F., Ross, A. G., McMillan, N. A., & Qadri, F. (2022). Toward typhoid fever elimination. *International Journal of Infectious Diseases*, *119*, 41-43.

Koch, R. (1884). An address on cholera and its bacillus. *British medical journal, 2*(1236), 453.

Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*(1), 1-21. https://doi.org/10.1207/s15327906mbr3001_1

Lachman, B. E., Resetar, S. A., Kalra, N., Schaefer, A. G., & Curtright, A. E. (2016). Water management, partnerships, rights, and market trends: An overview for Army installation managers (p. 0389). RAND.

Lanphear, B. P. (2005). Childhood lead poisoning prevention: too little, too late. Jama, 293(18), 2274-2276.

Lawinger H, Hlavsa M, Miko S, Kunz J, Thuneibat M, Gerdes, M, Gleason M, and Roberts, V. Waterborne Disease and Outbreak Surveillance System (WBDOSS) Summary Report, United States, 2021. Atlanta, Georgia: U.S. Department of Health and Human Services, CDC, 2023.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique, 15*(3), 209-225. https://doi.org/10.2307/3314912

Liang, K. Y., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health, 14*(1), 43-68. https://doi.org/10.1146/annurev.pu.14.050193.000355

Lin, L., St Clair, S., Gamble, G. D., Crowther, C. A., Dixon, L., Bloomfield, F. H., & Harding, J. E. (2023). Nitrate contamination in drinking water and adverse reproductive and birth outcomes: a systematic review and meta-analysis. Scientific Reports, 13(1), 563.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine, 19*(13), 1793-1819. doi: 10.1002/1097-0258(20000715)19:13<1793::aid-sim482>3.0.co;2-q.

Lyon, T. P., Montgomery, A. W., & Zhao, D. (2017). A change would do you good: Privatization, municipalization, and drinking water quality. In Academy of Management Proceedings (Vol. 2017, No. 1, p. 10499). Briarcliff Manor, NY 10510: Academy of Management.

Mahoney, L. E., Friedmann, C. T., Murray, R. A., Schulenburg, E. L., & Heidbreder, G. A. (1974). A waterborne gastroenteritis epidemic in Pico Rivera, California. *American Journal of Public Health, 64*(10), 963-968.

Maibach, E., Leiserowitz, A., Cobb, S., Shank, M., Cobb, K. M., & Gulledge, J. (2012). The legacy of climategate: Undermining or revitalizing climate science and policy?. *Wiley Interdisciplinary Reviews: Climate Change, 3*(3), 289-295.

Marill, K. A. (2004). Advanced statistics: Linear regression, part I: Simple linear regression. *Academic Emergency Medicine, 11*(1), 87-93. https://doi.org/10.1197/j.aem.2003.09.005

Masten, S. E. (2011). Public utility ownership in 19th-century America: The "aberrant" case of water. *The Journal of Law, Economics, & Organization, 27*(3), 604-654.

Masters, S., Wang, H., Pruden, A., & Edwards, M. A. (2015). Redox gradients in distribution systems influence water quality, corrosion, and microbial ecology. Water research, 68, 140-149.

McCullagh, P. (2019). *Generalized Linear Models* (2nd ed.). Routledge. https://doi.org/10.1201/9780203753736

McDonald, Y. J., & Jones, N. E. (2018). Drinking water violations and environmental justice in the United States, 2011–2015. *American Journal of Public Health, 108*(10), 1401-1407. https://doi.org/10.2105/AJPH.2018.304621

Medar, R., Rajpurohit, V. S., & Rashmi, B. (2017). Impact of training and testing data splits on accuracy of time series forecasting in machine learning. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICCUBEA.2017.8463779

Mohammadpour, E. (2013). A three-level multilevel analysis of Singaporean eighth-graders' science achievement. *Learning and Individual Differences, 26*, 212-220. https://doi.org/10.1016/j.lindif.2012.12.005

Monroe, R. G. (1906). The Gas, Electric Light, Water and Street Railway Services in New York City. The ANNALS of the American Academy of Political and Social Science, 27(1), 111-119.

Muñoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology, 48*(1), 1-33. https://doi.org/10.1177/00811750187779

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133-142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

National Oceanic and Atmospheric Administration. (2024). About our agency. Retrieved June 1, 2024, from https://www.noaa.gov/about-our-agency

NOAA National Centers for Environmental Information. (2024). Climate at a Glance: County

Mapping. National Oceanic and Atmospheric Administration. Retrieved June 1, 2024, from

https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping

NOAA National Centers for Environmental Information. (n.d.). About. National Oceanic and

Atmospheric Administration. Retrieved June 1, 2024, from https://ncei.noaa.gov/about-us

Navas-Acien, A., Guallar, E., Silbergeld, E. K., & Rothenberg, S. J. (2007). Lead exposure and

cardiovascular disease—a systematic review. Environmental health perspectives, 115(3), 472-

482.

Oberg, A. L., & Mahoney, D. W. (2007). Linear mixed effects models. In *Topics in Biostatistics*

(pp. 213-234). Springer. https://doi.org/10.1007/978-1-59745-530-5_11

Ohemeng, F. L., & Grant, J. K. (2011). Has the bubble finally burst? A comparative examination

of the failure of privatization of water services delivery in Atlanta (USA) and Hamilton

(Canada). *Journal of Comparative Policy Analysis: Research and Practice*, *13*(3), 287-306.

Osborne, J. W., & Waters, E. (2002). Multiple regression assumptions. *ERIC Digest*.

https://doi.org/10.7275/r222-hv23

Ostrom, V. (1962). The political economy of water development. The American Economic

Review, 52(2), 450-458.

Page, G. W. (1981). Comparison of groundwater and surface water for patterns and levels of

contamination by toxic substances. *Environmental Science & Technology, 15*(12), 1475-1481.

https://doi.org/10.1021/es00094a008

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*(1), 85-112. https://doi.org/10.1016/j.jsp.2009.09.002

Picard, R. R., & Berk, K. N. (1990). Data Splitting. The American Statistician, 44(2), 140–147. https://doi.org/10.1080/00031305.1990.10475704

Picetti, R., Deeney, M., Pastorino, S., Miller, M. R., Shah, A., Leon, D. A., ... & Green, R. (2022). Nitrate and nitrite contamination in drinking water and cancer risk: A systematic review with meta-analysis. Environmental Research, 210, 112988.

Ponkilainen, V. T., Uimonen, M., Raittio, L., Kuitunen, I., Eskelinen, A., & Reito, A. (2021). Multivariable models in orthopaedic research: A methodological review of covariate selection and causal relationships. *Osteoarthritis and Cartilage, 29*(7), 939-945. https://doi.org/10.1016/j.joca.2021.03.020

Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers, 52*, 145-158. https://doi.org/10.2307/621706

Powers, J. E., Mureithi, M., Mboya, J., Campolo, J., Swarthout, J. M., Pajka, J., Null, C., & Pickering, A. J. (2023). Effects of high temperature and heavy precipitation on drinking water quality and child hand contamination levels in rural Kenya. *Environmental Science & Technology, 57*(17), 6975-6988. https://doi.org/10.1021/acs.est.2c07284

Pyle, G. F. (1969). The diffusion of cholera in the United States in the nineteenth century. Geographical Analysis, 1(1), 59-75.

Rahman, T., Kohli, M., Megdal, S., Aradhyula, S., & Moxley, J. (2010). Determinants of environmental noncompliance by public water systems. *Contemporary Economic Policy, 28*(2), 264-274. https://doi.org/10.1111/j.1465-7287.2009.00150.x

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18*(4), 321-349. https://doi.org/10.2307/1165158

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage. https://doi.org/10.2466/pms.2002.94.2.671

Rights, J. D., & Sterba, S. K. (2023). On the Common but Problematic Specification of Conflated Random Slopes in Multilevel Models. *Multivariate Behavioral Research, 58*(6), 1106–1133. https://doi.org/10.1080/00273171.2023.2174490

Roberts, M. R., & Whited, T. M. (2013). Endogeneity in empirical corporate finance. In G. M. Constantinides, M. Harris, & R. M. Stulz (Eds.), *Handbook of the Economics of Finance* (Vol. 2, pp. 493-572). Elsevier. https://doi.org/10.1016/B978-0-44-453594-8.00007-0

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications, 15*(1), 1829. https://doi.org/10.1038/s41467-024-46150-w

Rosenfield, A. B., & Huston, R. (1950). Infant methemoglobinemia in Minnesota due to nitrates in well water. Minnesota medicine, 33(8), 789-796.

Rosenstock, T. S., Liptzin, D., Dzurella, K., Fryjoff-Hung, A., Hollander, A., Jensen, V., ... & Harter, T. (2014). Agriculture's contribution to nitrate contamination of Californian groundwater (1945–2005). Journal of Environmental Quality, 43(3), 895-907.

Roth, M. (1997). Cholera, community, and public health in gold rush Sacramento and San Francisco. Pacific Historical Review, 66(4), 527-551.

Roy, A. (2006). Estimating correlation coefficient between two variables with repeated observations using mixed effects model. *Biometrical Journal, 48*(2), 286-301. https://doi.org/10.1002/bimj.200510192

Royal, H., 't Mannetje, A., Hales, S., Douwes, J., Berry, M., & Chambers, T. (2024). Nitrate in drinking water and pregnancy outcomes: A narrative review of epidemiological evidence and proposed biological mechanisms. PLOS Water, 3(1), e0000214.

Sainani, K. L. (2014). Explanatory versus predictive modeling. *PM&R, 6*(9), 841-844. https://doi.org/10.1016/j.pmrj.2014.08.941

Salzman, J. E. (2022). The past, present and future of the safe drinking water act. UCLA School of Law, Public Law Research Paper, (22-21).

Sarkar, B. K. (2016). A case study on partitioning data for classification. *International Journal of Information and Decision Sciences, 8*(1), 73-91. https://doi.org/10.1504/IJIDS.2016.075788

Sawyer, W. A. (1916). The Typhoid Fever Death Rate in California. California State Journal of Medicine, 14(3), 110.

Sedlak, D. (2014). Water 4.0: The Past, Present, and Future of the World's Most Vital Resource. Yale University Press.

Sherris, A. R., Baiocchi, M., Fendorf, S., Luby, S. P., Yang, W., & Shaw, G. M. (2021). Nitrate in drinking water during pregnancy and spontaneous preterm birth: a retrospective within-mother analysis in California. Environmental health perspectives, 129(5), 057001.

Shmueli, G. (2010). To explain or to predict?. *Statistical Science, 25*(3), 289-310. https://doi.org/10.1214/10-STS330

Sigrist, F. (2022). Gaussian process boosting. *Journal of Machine Learning Research, 23*(232), 1-46. https://doi.org/10.48550/arXiv.2004.02653

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*(4), 323-355. https://doi.org/10.3102/10769986023004

Smith, A. H., Lopipero, P. A., Bates, M. N., & Steinmaus, C. M. (2002). Arsenic epidemiology and drinking water standards. Science, 296(5576), 2145-2146.

Smith, R. P., & Fuertes, A. M. (2010). Panel time series. In *The SAGE Handbook of Economic Methods* (pp. 239-258). SAGE. https://doi.org/10.1002/9781119504641.ch8

Snow, J. (1849). On the mode of communication of cholera.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles: SAGE.

Standiford, L. (2015). Water to the angels: William Mulholland, his monumental aqueduct, and the rise of Los Angeles. Ecco.

Stayner, L. T., Jensen, A. S., Schullehner, J., Coffman, V. R., Trabjerg, B. B., Olsen, J., ... & Sigsgaard, T. (2022). Nitrate in drinking water and risk of birth defects: Findings from a cohort study of over one million births in Denmark. The Lancet Regional Health–Europe, 14.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*, 307. https://doi.org/10.1186/1471-2105-9-307

Tomes, N. (2000). The making of a germ panic, then and now. American Journal of Public Health, 90(2), 191.

Triantafyllidou, S., & Edwards, M. (2012). Lead (Pb) in tap water and in blood: implications for lead exposure in the United States. Critical Reviews in Environmental Science and Technology, 42(13), 1297-1352.

Troesken, W. (1999). Typhoid rates and the public acquisition of private waterwork, 1880–1920. *The Journal of Economic History*, *59*(4), 927-948.

Tryland, I., Robertson, L., Blankenberg, A. G. B., Lindholm, M., Rohrlack, T., & Liltved, H. (2011). Impact of rainfall on microbial contamination of surface water. *International Journal of Climate Change Strategies and Management, 3*(4), 361-373. https://doi.org/10.1108/17568691111175650

Tulchinsky, T. H. (2018). John Snow, cholera, the broad street pump; waterborne diseases then and now. Case studies in public health, 77.

Turner, S. W., Rice, J. S., Nelson, K. D., Vernon, C. R., McManamay, R., Dickson, K., & Marston, L. (2021). Comparison of potential drinking water source contamination across one hundred US cities. Nature communications, 12(1), 7254.

Ulmer, R., & Gerlak, A. K. (2019). The Remunicipalization of Water Services in the United States. *Environment: Science and Policy for Sustainable Development*, *61*(4), 18–27. https://doi.org/10.1080/00139157.2019.1615826

United States Census Bureau. (1850). Seventh Census of the United States, 1850. Retrieved May 22, 2024, from https://www2.census.gov/library/publications/decennial/1850/1850a/1850a-47.pdf

United States Census Bureau. (2023-a). Annual estimates of the resident population for incorporated places in California: April 1, 2020 to July 1, 2022. Retrieved May 22, 2024, from https://www2.census.gov/programs-surveys/popest/tables/2020-2022/cities/totals/SUB-IP-EST2022-POP-06.xlsx

United States Census Bureau. (2023-b). About the Bureau. Retrieved June 1, 2024, from https://www.census.gov/about.html

United States Census Bureau. (2024). American Community Survey (ACS). Retrieved June 1, 2024, from https://www.census.gov/programs-surveys/acs

United States Census Bureau. (n.d.). Explore Census Data. Census.gov. Retrieved June 1, 2024, from https://data.census.gov

United States Congress. Senate Committee on Environment and Public Works. (1986). The Safe Drinking Water Act as amended by the Safe Drinking Water Act amendments of 1986: (Public law 99-339, June 19, 1986).

United States Congress. Senate Committee on Environment and Public Works. (1996). The Safe Drinking Water Act as amended by the Safe Drinking Water Act of 1996, public law 104-182, August 6, 1996.

United States Environmental Protection Agency. (2001). Drinking Water Standard for Arsenic. https://nepis.epa.gov/Exe/ZyPdf.cgi?Dockey=20001XXC.txt

United States Environmental Protection Agency. (2021). Lead and Copper Rule Revisions. https://www.epa.gov/system/files/documents/2024-04/revised-508_lcrr-compliance-fact-sheet_4.17.24.pdf#page=1.00

United States Environmental Protection Agency. (2022). SDWIS Federal Reports Advanced Search. EPA. https://sdwis.epa.gov/ords/sfdw/sdwis_fed_reports_public/1

United States Environmental Protection Agency (2023-a). *Drinking water regulations*. Retrieved June 14, 2024, from https://www.epa.gov/dwreginfo/drinking-water-regulations

United States Environmental Protection Agency. (2023-b). Information about public water systems. Retrieved June 1, 2024, from https://www.epa.gov/dwreginfo/information-about-public-water-systems

United States Environmental Protection Agency. (2024). Our mission and what we do. Retrieved June 1, 2024, from https://www.epa.gov/aboutepa/our-mission-and-what-we-do

Van Oosten, R. (2016). The Dutch Great Stink: The end of the cesspit era in the pre-industrial towns of Leiden and Haarlem. European Journal of Archaeology, 19(4), 704-727.

VanDerslice, J. (2011). Drinking water infrastructure and environmental disparities: Evidence and methodological considerations. *American Journal of Public Health, 101*(S1), S109-S114. https://doi.org/10.2105/AJPH.2011.300189

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data?. *Ecology, 88*(11), 2766-2772. https://doi.org/10.1890/07-0043.1

Warner, M. E. (2021). Key issues in water privatization and remunicipalization. Utilities Policy, 73, 101300.

Watson, R. (2015). Quantitative research. *Nursing Standard: Official Newspaper of the Royal College of Nursing, 29*(31), 44-48. https://doi.org/10.7748/ns.29.31.44.e8681

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2019). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation, 18*(1), 11. https://doi.org/10.7275/55hn-wk47

Yaremych, H. E., Preacher, K. J., & Hedeker, D. (2021). Centering categorical predictors in multilevel models: Best practices and interpretation. Psychological Methods. https://doi.org/10.1037/met0000434

Zernike, K. (2003, May 2). Arsenic case is considered homicide, Maine police say. The New York Times. https://www.nytimes.com/2003/05/02/us/arsenic-case-is-considered-homicide-maine-police-say.html

Zhang, T., & Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *Annals of Statistics, 33*(4), 1538-1579. https://doi.org/10.1214/009053605000000255

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R* (Vol. 574). Springer. https://doi.org/10.1007/978-0-387-87458-6

Appendix

Figure 20

Estimates of the 1st RIFS Model in Stage 2

Dot-and-Whisker Plot of CWC-1 IRRs
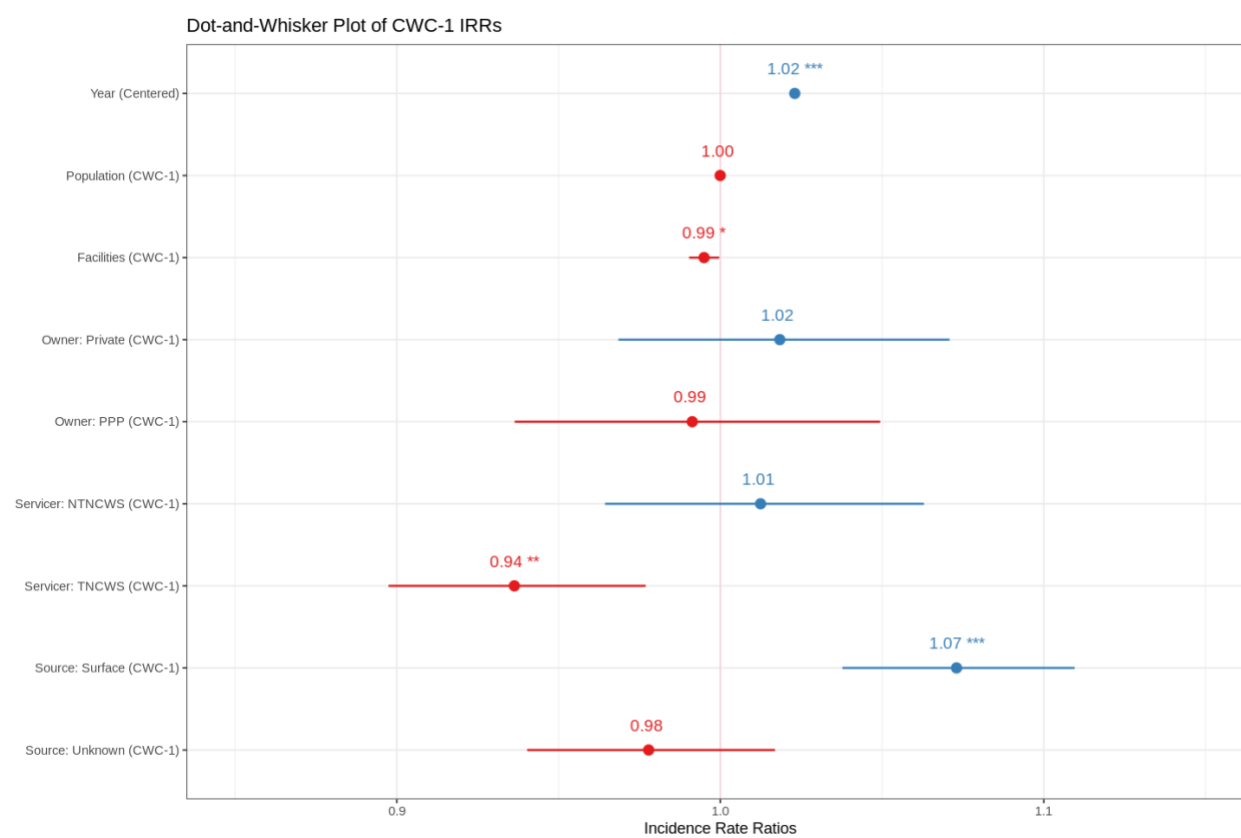
Figure 21

Estimates of the 2nd RIFS Model in Stage 2

Dot-and-Whisker Plot of CWC-1/CGM-2 IRRs
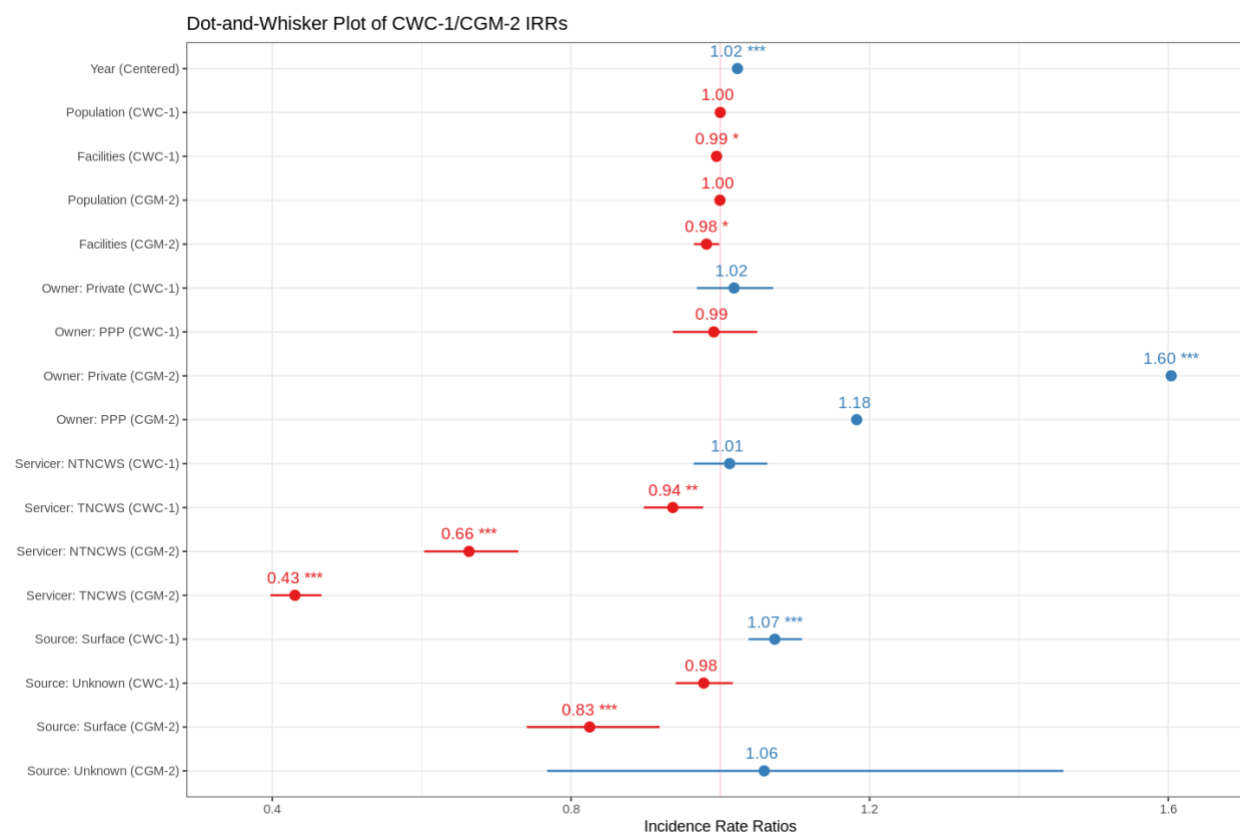
Figure 22

Estimates of the 3rd RIFS Model in Stage 2

Dot-and-Whisker Plot of CWC-1/CWC-2 IRRs
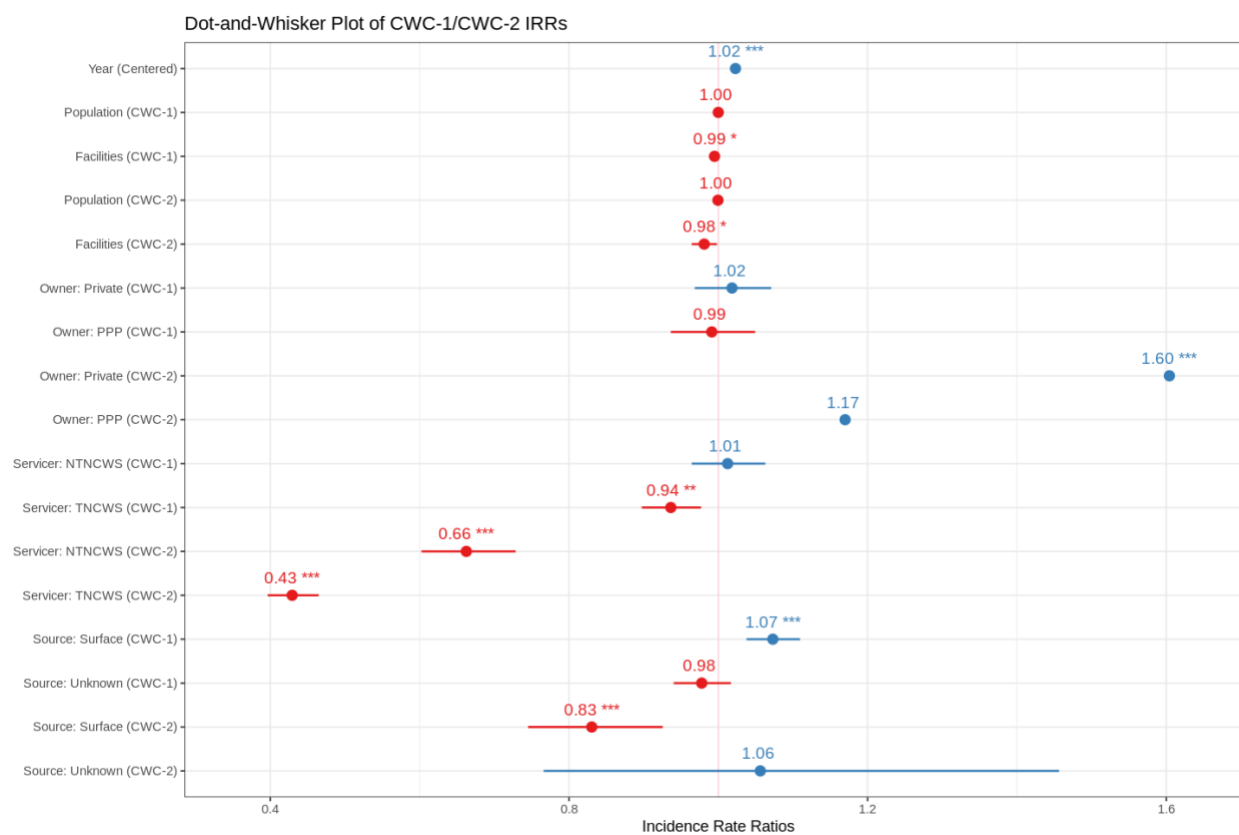
Figure 23

Comparison of Stage-2 Model Estimates

Dot-and-Whisker Plots of RIFS Models
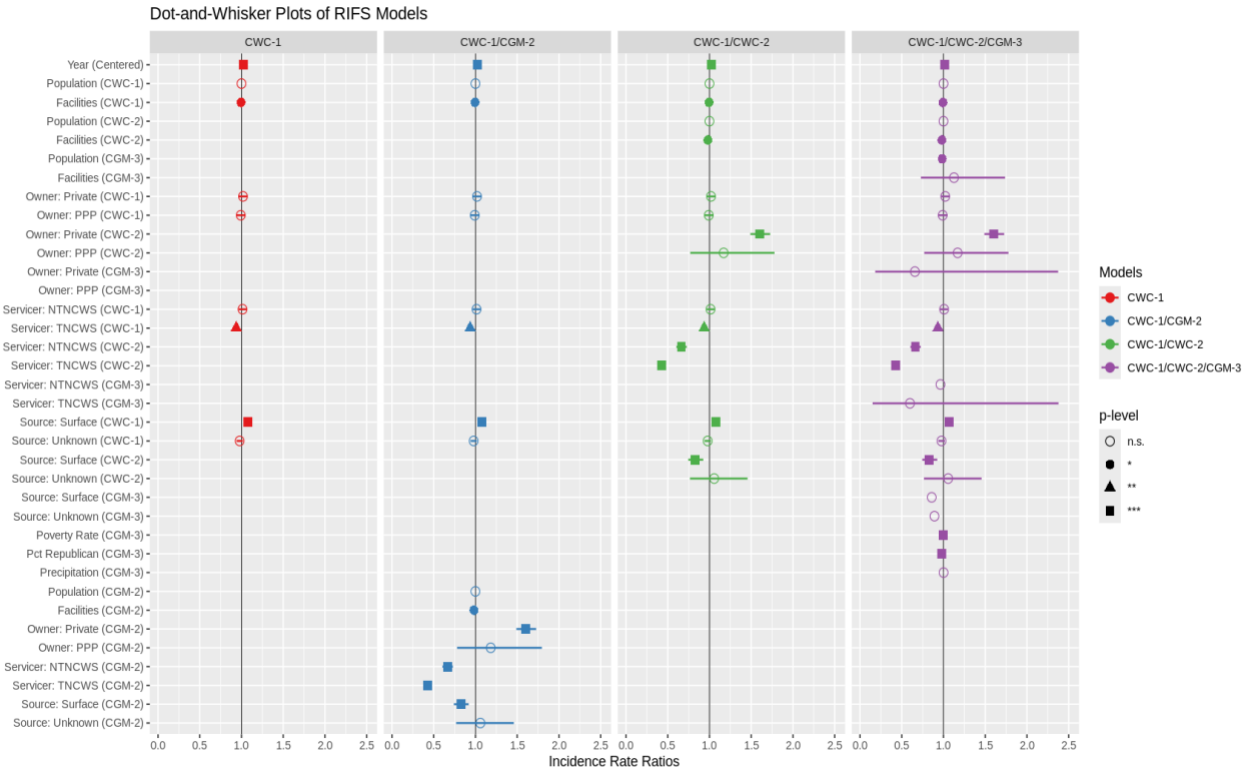
Figure 24

Estimates of the 1st SLI Model in Stage 4

Dot-and-Whisker Plot of SLI IRRs
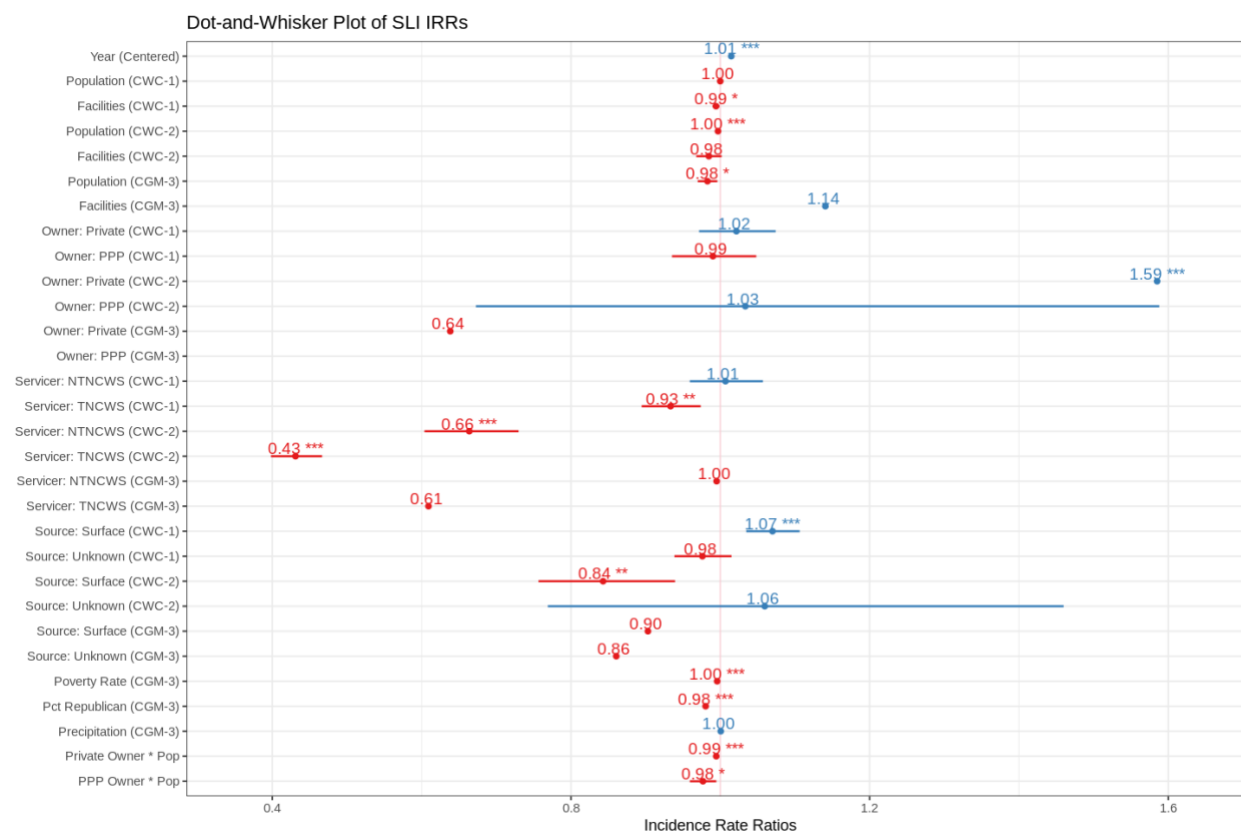
Figure 25

Estimates of the 2nd SLI Model in Stage 4

Dot-and-Whisker Plot of SLI IRRs
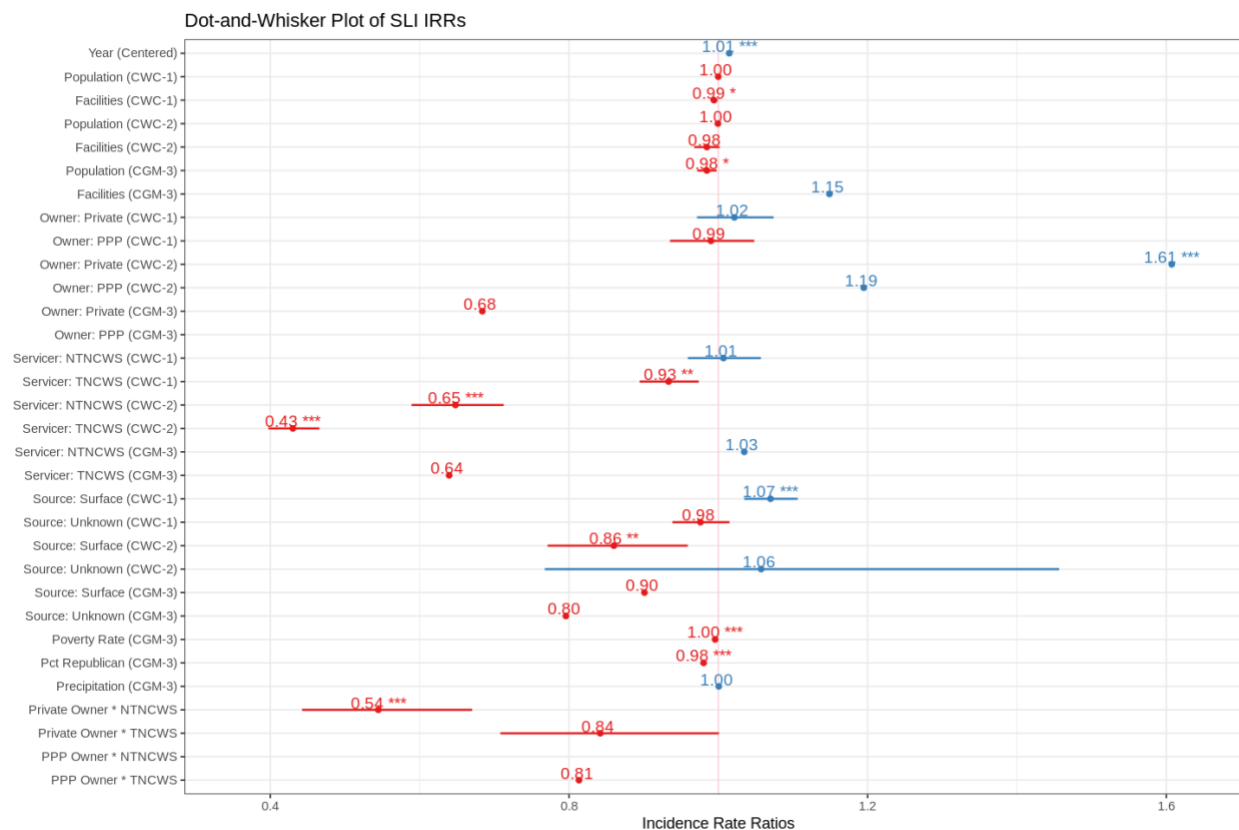
Figure 26

Estimates of the 3rd SLI Model in Stage 4

Dot-and-Whisker Plot of SLI IRRs
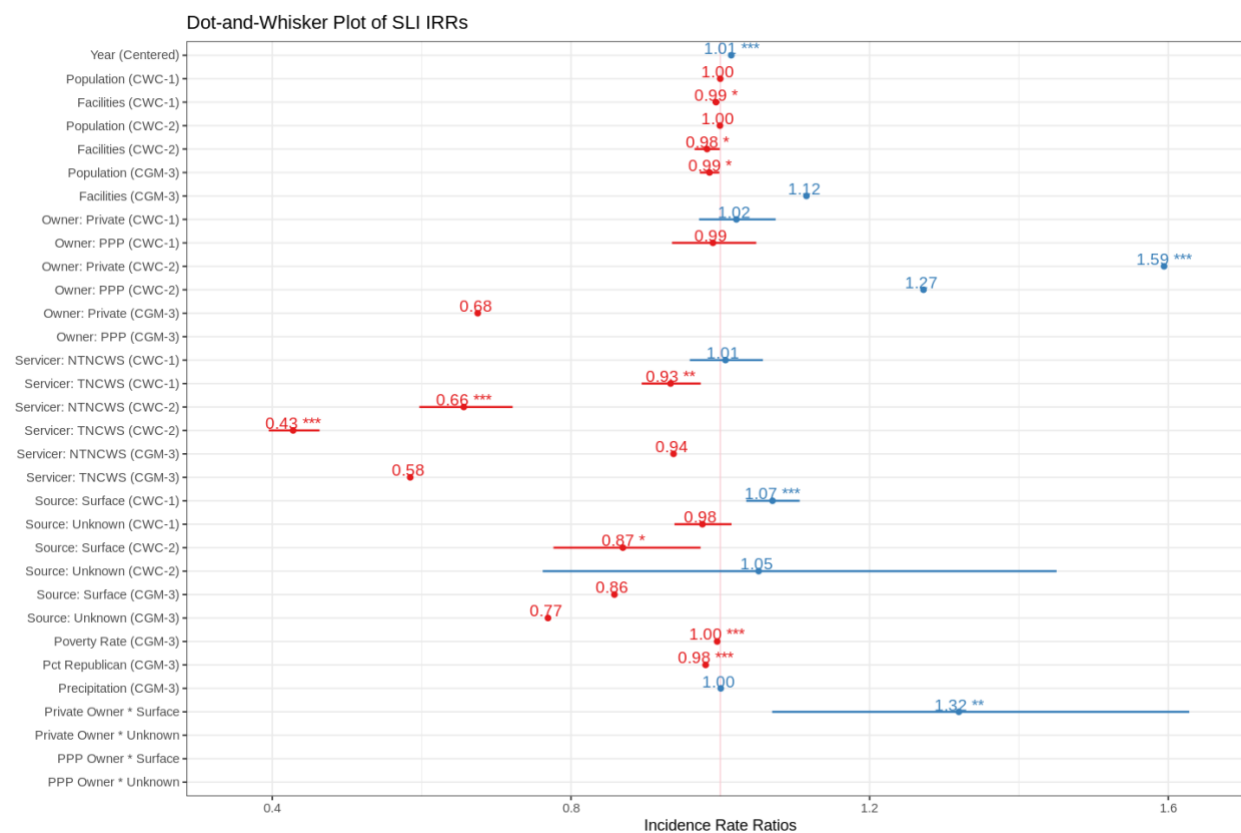
Figure 27

Estimates of the 1st CLI Model in Stage 5

Dot-and-Whisker Plot of CLI IRRs
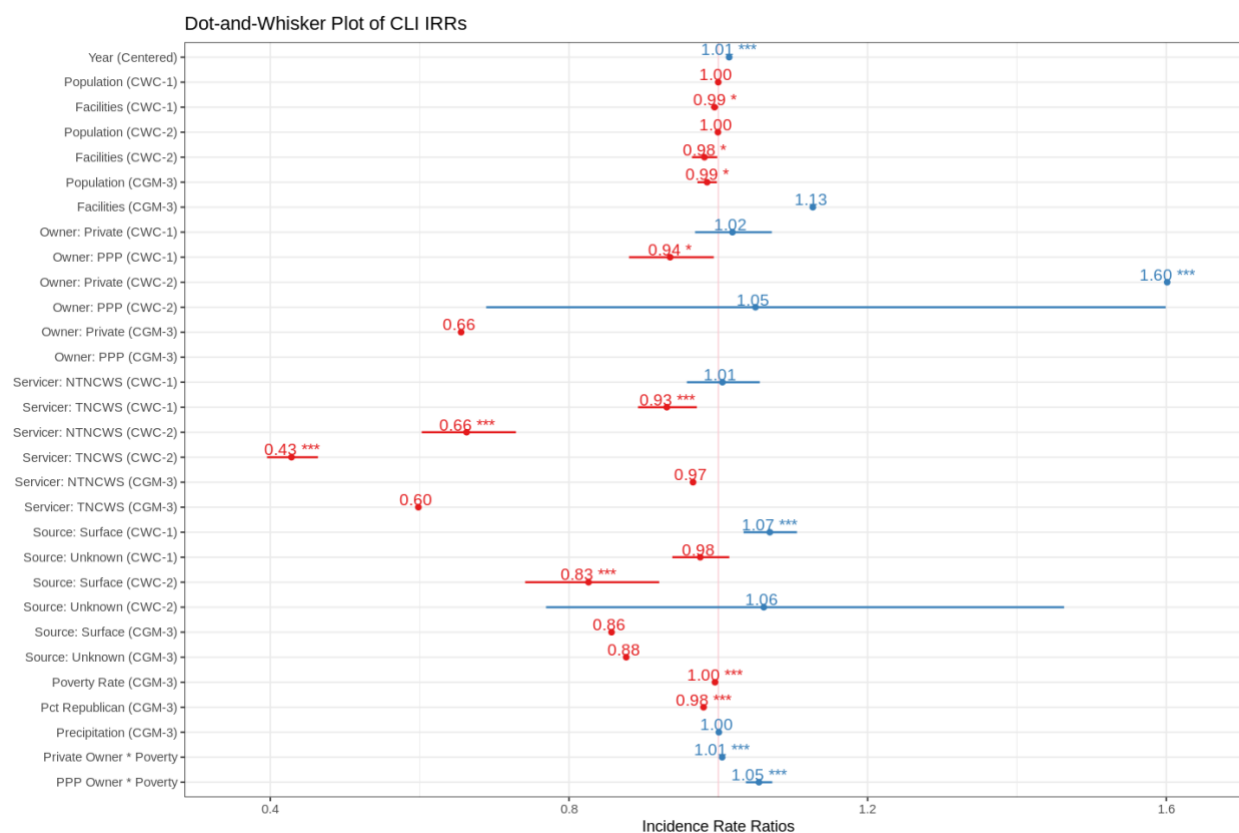
Figure 28

Estimates of the 2nd CLI Model in Stage 5

Dot-and-Whisker Plot of CLI IRRs