

# NATIONAL ASSESSMENT OF U.S. COMMUNITY WATER SYSTEMS

## Charges, Health Violations, and Financial Performance

*Insights from Linear and Nonlinear Statistical Analyses Across Public and Private Utilities*

---

Prepared for the  
**Mississippi River Cities and Towns Initiative (MRCTI)**  
Grant # PG013603

By

 COLUMBIA CLIMATE SCHOOL  
COLUMBIA WATER CENTER

Authored by  
Upmanu Lall and Mengling Qiao, Columbia Water Center  
and  
Gabriel Watson and EmmaLi Tsai, Environmental Policy Innovation Center (EPIC)

2026

# Executive Summary

**Scope and Purpose.** This report presents findings from a comprehensive national study of United States community water systems (CWSs), sponsored by the Mississippi River Cities and Towns Initiative (MRCTI). The study was designed to: (a) assemble an integrated database of water utility characteristics, health and total Safe Drinking Water Act (SDWA) violations, financial data, and customer water charges; (b) examine how these performance measures relate to community demographics, water source type, system scale, utility ownership, and state context; and (c) provide a national reference for interpreting the position of Mississippi River corridor utilities. The study covers 43,532 utilities in the full sample (EMMA dataset), with nested subsets of 8,007 public utilities with municipal financial data and 1,235 public water system IDs across 242 private companies. *Generalized Linear Models (GLM)* and *Random Forest (RF)* machine learning methods were applied, providing complementary linear and nonlinear perspectives from the same data. The financial data used is for 2022. This is the largest analysis to date of the factors associated with the financial and violation metrics of US water utilities.

## Mississippi River Corridor — MRCTI Sub-Report

The Mississippi River corridor analysis for the ten corridor states (LA, MN, WI, IA, IL, MO, KY, AR, TN, MS) is presented in a companion sub-report. Key corridor findings are noted in sidebars throughout this report.

## Part A — For Water Utility Managers and Policy Audiences

### Five findings that should matter to every utility manager

- **Ownership matters for charges, not for violations.** After controlling for state, source, demographics, and financial structure, publicly owned utilities charge about \$41 more per customer per year than privately owned utilities. However, ownership has almost no independent effect on health violations. What drives violations is where a utility operates, its water source, and, for publicly owned utilities, the poverty level and financial structure of its community. Conditional on all other variables being held constant, the marginal dependence of average water charges and health violations on PCD2 (Wealthier, Urban Density & Rental) shows a distinct performance advantage for private utilities over public utilities, in that they appear to charge lower rates and have fewer violations across all demographic groups. Charges increase, and violations decrease as urban wealth rises up to a point, then remain constant.
- **States set the context more than any other single factor.** State-level effects explain 20% of the variation in average charges and 7–11% of health violations, larger than any other predictor group. After controlling for everything else, Oklahoma and Louisiana public utilities have more health violations than the national average. Notably, the Oklahoma effect is concentrated in Public utilities – its private utilities do not have elevated violation rates. Public utilities in Washington, California, and Oregon charge substantially more than those in other states. Public utilities in Mississippi and Louisiana charge substantially less. These patterns persist after controlling for water source, demographics, and finances, implying that regulatory frameworks, geological environments, and infrastructure histories create durable state-level environments that determine outcomes rather than other measured factors.
- **Public utilities serving the poorest communities charge more, not less.** This counterintuitive finding, confirmed by both the GLM and the RF, is one of the most important equity findings in the study. In the public utility dataset, the poverty gradient is the strongest demographic predictor of charges, and the relationship is continuous and accelerating: the more poverty-concentrated the service area, the higher the average charge, all else equal. The general trend is for charges to increase with economic affluence for both public and private utilities. However, accounting for other

factors, charges for public utilities serving small communities increase with poverty. The same communities face higher charges AND more violations, resulting in a compounding disadvantage. Overall, the highest water charges in the nation seem to be associated with suburban, middle-income communities with high homeownership and relatively homogeneous demographics. Utilities with a high standard deviation of charges within the utility, possibly reflecting a high variation in use characteristics, housing density and type, or subsidy structures, tend to have higher average charges.

- **Small systems are the highest-risk archetype in both sectors.** Private utilities serving fewer than 10,000 people frequently operate at or near a financial deficit. Public utilities face similar stress below 1,000 people. Both GLM and RF confirm that larger systems have fewer violations and better cost recovery than smaller systems.
- **Financial stress links to worse health violations, but the relationship is not simple.** Operating expense pressure is associated with more violations. However, the RF models reveal a nuance: for public utilities, violations are highest not just when revenue is severely deficient but also when utilities' revenue far exceeds their operating expenses, a U-shaped relationship around the breakeven point. The goal of rate-setting should be cost recovery with reinvestment, not surplus extraction. High Taxes relative to revenue lead to financial stress for private utilities and contribute to higher health violations where revenue-to-expense margins are weaker. Pennsylvania and California have regulated investor-owned utilities, some of which face these pressures. Public utility violations decrease with Economic Affluence, with the steepest decline occurring in the poverty tail, and a flat plateau at above-average income levels, suggesting that economic resources above the poverty threshold are sufficient to eliminate the income-driven compliance deficit.
- **Utilities with groundwater as the source tend to have lower charges, expenses, and violations than those with surface water as the source, all other factors being equal.** These factors interact strongly with demographic factors and ownership. High poverty and reliance on surface water sources interact to amplify costs, financial stress, and violations. Specifically for the public utility subsample, surface water systems average 3.4 times more violations than groundwater systems

### What this study does differently

**Survey-based charges rather than rate formula estimates.** Most prior research estimates customers' payments by applying a utility's published rate schedule to a hypothetical household with a fixed water volume. This study instead draws average water charges and charge variability (the spread of what customers actually pay, technically the standard deviation of annual water charges) directly from the 2022 American Community Survey, capturing the real heterogeneity of bills paid by actual households in each utility's service area. This means our charge estimates reflect affordability as households actually experience it, including the effects of tiered rates, multi-family versus single-family service, and local consumption patterns.

**Two analytical lenses.** GLMs identify which factors are statistically associated with charges and violations and estimate the direction and magnitude of these associations. Random Forest models reveal the shape of those relationships, where thresholds exist, where two factors multiply each other's effects, and where relationships reverse at extreme values. In every case where the two methods are directly comparable, their directional findings agree. The RF consistently reveals that underlying relationships are nonlinear, something GLMs cannot detect by construction. This allows nuances across small utilities, different income groups, and interactions across water source, demographic and financial covariates to be detected.

### Key operational recommendations

- **Identify compound-disadvantage utilities.** Utilities serving communities with high poverty AND high violations AND high charges are the policy emergency. Federal programs that address rate affordability in isolation from compliance infrastructure will not break this cycle. See Appendix F for

access to a dashboard that facilitates the identification of such utilities. Health violations and affordability are not independent problems amenable to independent solutions: they are co-determined by the same infrastructure economics, demographic context, and governance dynamics. Policy frameworks that address affordability through rate suppression without investing in compliance will not break the cycle of compounding disadvantage; frameworks that address compliance without affordability assistance will impose unaffordable burdens on the communities most in need of infrastructure improvements

- **Debt leverage is the largest financial driver of public charges — and it is addressable.** The RF model shows a threshold effect: utilities above a specific debt-to-revenue threshold charge approximately \$46/yr more. Grant rather than loan financing for the most highly leveraged utilities would reduce both charges and financial fragility and thus may affect violations.
- **Water source type is the dominant driver of violations in public utilities.** Surface water utilities average 3.4 times as many health violations as groundwater utilities, and the difference reaches 7× in high-poverty service areas. Infrastructure investment should prioritize source type × poverty level as an eligibility criterion. The caveat is that all these analyses assume that violations are actually being reported.
- **Prioritize Investment in Performance and Equity across community scale, not just Affordability.** Larger communities can achieve affordability through cross-subsidies and tiered pricing structures. Smaller, poorer communities struggle because they lack access to such options. Large, investor-owned utilities can potentially achieve a financial balance across their assets – an argument made for the consolidation of smaller public utilities; however, they are still constrained by resources and regulatory pressures at the community water system level. Investment vehicles that cover the smaller and poorer communities as a group, regardless of ownership, and new designs for economical, distributed water and wastewater systems that use groundwater and consider localized wastewater treatment and reuse that reduce conveyance and operational costs, while assuring point of use quality and service reliability, emerge as a possible strategy for the large number of underserved, smaller water systems. The observation that larger systems have better cost-effectiveness and lower violations is tied to their financial base, ability to raise capital (debt profile), and cross-subsidization. We have to think of ways this can be achieved for smaller systems collectively at the state level.
- **For Mississippi River corridor utilities:** see the companion MRCTI sub-report for state-by-state positioning relative to this national analysis.

## Part B — For Technical and Academic Audiences

### Study design and methodological innovations

This study assembles a three-tier data structure balancing breadth with financial depth. The EMMA dataset (N=43,532) integrates SDWIS violation records, EPA-EPIC geospatial mapping, and 2022 ACS demographic data for all community water systems nationally. The Public dataset (N=8,007) enriches the EMMA subset with 2022 municipal finance data. The Private dataset (N=1,235 PWSIDs / 242 companies) integrates 2022 Bluefield and NAWC annual financial reports aggregated at the company level. Principal Component Analysis was applied separately within each dataset and variable group to produce uncorrelated predictor sets.

Two complementary methods were applied: Lasso-regularized GLMs (Gaussian for charges, Poisson for violation counts, Logistic for operating deficit) and Random Forest models with SHAP decomposition, 1D and 2D partial dependence plots, and Friedman H-statistics for interaction quantification. RF models consistently achieve higher variance explained (Cross Validated  $R^2$  is 0.20–0.40 for violations; 0.40+ for charges) and reveal nonlinear functional forms that provide insights beyond linear regression.

## Model performance summary

**Table 1.** Model performance: GLM ( $R^2$ ) and RF (cross-validated  $R^2$ ) for primary models. RF models for 5-year health violations, 10-year total violations, and charge variability are consistent in direction; results available from authors.

| Dataset | Response           | GLM $R^2$ | RF train $R^2$ | RF CV $R^2$ | Dominant GLM predictor | Dominant RF predictor                   |
|---------|--------------------|-----------|----------------|-------------|------------------------|-----------------------------------------|
| EMMA    | Avg. charge        | 0.23      | 0.58           | 0.40        | State (19.9%)          | PCD2 Urban Wealth, state_CA             |
| EMMA    | 10-yr health viol. | 0.11      | 0.36           | 0.20        | State (7.3%)           | state_OK, source_GW                     |
| Public  | Avg. charge        | 0.62      | ~0.65          | ~0.45       | Charge SDev + State    | state_WA , Economic Affluence (PCD1)    |
| Public  | 10-yr health viol. | 0.17      | ~0.45          | ~0.25       | State (10.9%)          | state_OK, water source                  |
| Private | Avg. charge        | 0.63      | 0.42           | 0.24        | Demographics (31%)     | state_CA, PCD3 Knowledge Eco.           |
| Private | 10-yr health viol. | 0.25      | 0.19           | 0.06        | Financial structure    | PCF2 Tax Impact<br>PCF1 Financial Scale |

## Cross-cutting technical findings

- **State effects are larger than all other factors combined** for charges in EMMA and public models. State partial  $R^2$  (GLM) reaches 19.9% for EMMA charges and 10.9% for public violations, exceeding demographics, financial structure, ownership, and source type individually. RF SHAP maps confirm large, geographically structured state premiums that persist after controlling for covariate sets.
- **Nonlinearity is the rule, not the exception.** Every significant predictor in the RF models exhibits threshold, sigmoid, step-function, or U-shaped behavior. The most consequential nonlinearities are: (i) poverty-charge acceleration in public utilities; (ii) source  $\times$  poverty multiplicative interaction producing 1.0 to 7.0x violations; (iii) PCF2's binary step controlling both private charges (+\$20.5) and violations (+2.34) under high tax burdens; (iv) debt leverage's binary \$46 premium in public charges.
- **Compound disadvantage is confirmed across all three datasets.** Poor communities served by public utilities pay more AND have more violations. The PCD1 (Economic Affluence)  $\times$  PCD3 (high rates of health uninsurance) superadditive interaction in public charges (poor + uninsured communities pay ~\$25 above additive prediction) identifies the highest-priority compound-disadvantage communities.

## Contents

|                                                                                                              |    |
|--------------------------------------------------------------------------------------------------------------|----|
| Executive Summary.....                                                                                       | 2  |
| Part A — For Water Utility Managers and Policy Audiences .....                                               | 2  |
| Five findings that should matter to every utility manager .....                                              | 2  |
| What this study does differently .....                                                                       | 3  |
| Key operational recommendations.....                                                                         | 3  |
| Part B — For Technical and Academic Audiences .....                                                          | 4  |
| Study design and methodological innovations .....                                                            | 4  |
| Model performance summary.....                                                                               | 5  |
| Cross-cutting technical findings.....                                                                        | 5  |
| 1. Introduction.....                                                                                         | 11 |
| 1.1 Motivation and Sponsorship.....                                                                          | 11 |
| 1.2 Research Questions .....                                                                                 | 11 |
| 1.3 A Methodological Note: Survey-Based Charges.....                                                         | 11 |
| 1.4 Report Organization.....                                                                                 | 12 |
| 2. Performance of U.S. Water Utilities: Health Violations, Charges, and Finances — A Decade of Evidence..... | 13 |
| 2.1 Introduction .....                                                                                       | 13 |
| 2.2 National Trends and Geographic Patterns in Health-Based Violations .....                                 | 14 |
| 2.2.1 Long-Run National Patterns .....                                                                       | 14 |
| 2.2.2 Geographic Concentration and State-Level Effects .....                                                 | 15 |
| 2.3 Public versus Private Ownership: A Contested Empirical Question .....                                    | 15 |
| 2.3.1. The State of the Evidence .....                                                                       | 15 |
| 2.3.2. Ownership and Charges: A Different Story .....                                                        | 16 |
| 2.4. Demographic Drivers of Disparate Outcomes .....                                                         | 17 |
| 2.4.1. Race, Ethnicity, and Health-Based Violations .....                                                    | 17 |
| 2.4.2. Income, Poverty, and the Affordability Paradox .....                                                  | 17 |
| 2.4.3. Rural Status and System Scale.....                                                                    | 18 |
| 2.5. Financial Determinants and Infrastructure Investment .....                                              | 19 |
| 2.5.1. Debt, Revenue Recovery, and Public Utility Charges .....                                              | 19 |
| 2.5.2. Consolidation as a Policy Response .....                                                              | 19 |
| 2.6. Synthesis and Discussion .....                                                                          | 20 |
| 2.6.1. What the Literature Establishes with Confidence.....                                                  | 20 |
| 2.6.2. Areas of Persistent Uncertainty .....                                                                 | 20 |
| 2.6.3. A Compound Disadvantage Framework .....                                                               | 21 |
| 2.7 Conclusions and Research Priorities .....                                                                | 21 |
| 2.8 What This Study Adds .....                                                                               | 22 |
| 2.9 References.....                                                                                          | 22 |
| 3. Data Sources, Datasets, and Analytical Methods .....                                                      | 26 |

|                                                                                             |    |
|---------------------------------------------------------------------------------------------|----|
| 3.1 Data Sources .....                                                                      | 26 |
| 3.2 Three-Tier Dataset Structure .....                                                      | 27 |
| 3.3 Response Variables and Modeling Structure.....                                          | 27 |
| 3.4 Principal Component Analysis .....                                                      | 28 |
| 3.5 Generalized Linear Models with Lasso Regularization .....                               | 28 |
| 3.6 Random Forest Models.....                                                               | 29 |
| 4. Exploratory Analysis .....                                                               | 30 |
| 4.1 Who Serves Whom? The Structural Context of Ownership.....                               | 30 |
| 4.2 Water Charges: Similar Medians, Strong Context Effects .....                            | 32 |
| 4.3 Health Violations: Ownership, Source, and Geography.....                                | 36 |
| 4.4 Financial Performance: Cost Recovery and the Role of Taxes .....                        | 37 |
| .....                                                                                       | 39 |
| 5. Statistical Inference: GLM and Random Forest Results.....                                | 40 |
| 5.1 EMMA Dataset: Full Utility Population without Financial Data.....                       | 40 |
| 5.1.1 Average Water Charges & 10 year Health Violations— EMMA.....                          | 41 |
| 5.1.2 10-Year Health Violations — EMMA .....                                                | 42 |
| 5.2 Public Utilities.....                                                                   | 43 |
| 5.2.1 Average Water Charges — Public .....                                                  | 43 |
| 5.2.2 10-Year Health Violations — Public .....                                              | 44 |
| 5.3 Private Utilities .....                                                                 | 44 |
| 5.3.1 Average Water Charges — Private .....                                                 | 44 |
| 5.3.2 10-Year Health Violations — Private.....                                              | 45 |
| 5.4 Operating Deficit (Financial Performance) Models .....                                  | 46 |
| 5.5 Public vs Private Assessment across response variables for significant covariates ..... | 46 |
| 6. Synthesis .....                                                                          | 48 |
| Part A — Synthesis for Water Utility Managers and Policy Audiences.....                     | 48 |
| S.A.1 Context matters more than ownership.....                                              | 48 |
| S.A.2 The compound disadvantage problem .....                                               | 48 |
| S.A.3 The small system problem is structural, not accidental.....                           | 49 |
| S.A.4 Water source type: the overlooked compliance driver.....                              | 49 |
| S.A.5 The Deep South underpricing trap .....                                                | 49 |
| Part B — Technical and Academic Synthesis.....                                              | 50 |
| S.B.1 Where linear and nonlinear methods agree .....                                        | 50 |
| S.B.2 What nonlinear analysis reveals that linear models miss .....                         | 50 |
| S.B.3 The ownership contrast: structural rather than governance .....                       | 50 |
| S.B.4 Confirming the literature — and where this study diverges .....                       | 51 |
| S.B.5 Research priorities.....                                                              | 51 |
| References .....                                                                            | 52 |
| Acknowledgements:.....                                                                      | 52 |

|                                                                                      |    |
|--------------------------------------------------------------------------------------|----|
| Appendix A: Database Development.....                                                | 53 |
| A.1 Overview.....                                                                    | 53 |
| A.2 Bluefield Research Financial Reports .....                                       | 53 |
| A.3 NAWC Financial Data.....                                                         | 53 |
| A.4 Private Financial Database.....                                                  | 53 |
| A.5 SDWIS.....                                                                       | 53 |
| A.6 American Community Survey (ACS) .....                                            | 53 |
| A.7 Municipal Finance Data .....                                                     | 53 |
| Appendix B: Principal Component Analysis Methodology and Component Definitions ..... | 55 |
| B.1 Rationale and Approach.....                                                      | 55 |
| B.2 EMMA Demographic Components .....                                                | 55 |
| B.3 Public Utility Demographic Components .....                                      | 55 |
| B.4 Private Utility Demographic Components (Selected) .....                          | 56 |
| B.5 Public Utility Financial Components .....                                        | 56 |
| B.6 Private Utility Financial Components.....                                        | 57 |
| B.7 Utility Scale Components .....                                                   | 57 |
| Appendix C: Statistical Methods Reference .....                                      | 58 |
| C.1. Generalized Linear Models (GLM) with Lasso: Mathematical Formulation.....       | 58 |
| C.1.1 Lasso Regularization .....                                                     | 59 |
| C.1.2. Cross-Validation for Model Selection.....                                     | 60 |
| C.2. Analytical Framework: Random Forest Regression.....                             | 61 |
| C.2.1 Overview of Random Forests .....                                               | 61 |
| C.2.2 Interpreting Feature Importance .....                                          | 61 |
| C.2.3 SHAP Analysis — Direction and Magnitude of Effects.....                        | 62 |
| C.2.4 Partial Dependence Plots — Marginal Response Functions.....                    | 63 |
| C.2.6 Model Limitations and Interpretive Caveats.....                                | 64 |
| C.3 Methodological Limitations .....                                                 | 64 |
| C.3.1. Observational Design and Causal Inference .....                               | 64 |
| C.3.2 Model Specification in GLM.....                                                | 65 |
| C.3.3 Regularization Bias (Lasso).....                                               | 65 |
| C.3.4 Cross-Validation and Model Selection Uncertainty.....                          | 65 |
| C.3.5 Random Forest Interpretability Constraints .....                               | 66 |
| C.3.6 Heterogeneity Across States and Regulatory Environments.....                   | 66 |
| C.3.7 Measurement and Reporting Error .....                                          | 66 |
| C.3.8 External Validity .....                                                        | 66 |
| C.3.9 Predictive vs. Structural Modeling .....                                       | 67 |
| C.3.10 Bias-Variance Considerations and Limitations .....                            | 67 |
| C.3.11 Summary .....                                                                 | 68 |
| Appendix D: Detailed GLM Results and Discussion.....                                 | 69 |

|                                                                      |     |
|----------------------------------------------------------------------|-----|
| D.1 Executive summary .....                                          | 69  |
| Pricing level: average customer charges .....                        | 69  |
| Pricing heterogeneity: variation within a utility.....               | 69  |
| Financial stress: operating expenses versus revenues.....            | 69  |
| Regulatory performance: health and total violations .....            | 69  |
| Public versus private: conceptual contrasts .....                    | 70  |
| D.2 A guide to this appendix: .....                                  | 71  |
| D.3 Emma: .....                                                      | 72  |
| D.4 Public:.....                                                     | 75  |
| D.4.1 Executive Summary.....                                         | 75  |
| D.5 Private:.....                                                    | 84  |
| D6. Overall Comparative Conclusions at a Glance: .....               | 88  |
| D7: Technical Influence Matrix .....                                 | 89  |
| D8. Amplifiers vs. Muters .....                                      | 89  |
| D8 Geographic and Demographic Structure of Utility Performance ..... | 90  |
| D9 Technical appendix (GLM results) .....                            | 93  |
| Appendix E Random Forest Models.....                                 | 96  |
| E1. Emma .....                                                       | 96  |
| E1.1 Average Water Charge.....                                       | 97  |
| E1.1.1 Synthesis: Conceptual Model of Water Charge Drivers.....      | 102 |
| Tier 1: Primary Continuous Drivers (Socioeconomic Character).....    | 102 |
| Tier 2: Geographic Regulatory Premium/Discount.....                  | 103 |
| Tier 3: Secondary Demographic Modifiers .....                        | 103 |
| Tier 4: Complex/Interaction-Dependent Features .....                 | 103 |
| Tier 5: Sparse/Threshold Effects.....                                | 103 |
| E1.1.2 Conclusions .....                                             | 103 |
| E1.2 10-year Health Violations .....                                 | 104 |
| <i>Water Source Effects</i> .....                                    | 111 |
| Key Contrasts with the Water Charges Model .....                     | 113 |
| Key Nonlinearities and Their Substantive Implications .....          | 113 |
| E2. Public.....                                                      | 115 |
| E2.1Average Water Charges .....                                      | 115 |
| .....                                                                | 115 |
| Feature Importance .....                                             | 117 |
| Interaction Strengths (Friedman H-Statistic) .....                   | 119 |
| Partial Dependence Analysis .....                                    | 120 |
| E2.2 10 year Health Violations.....                                  | 126 |
| Feature Importance Overview .....                                    | 128 |
| Interaction Strengths (Friedman H-Statistic) .....                   | 130 |

|                                                                               |     |
|-------------------------------------------------------------------------------|-----|
| Marginal Effect of Water Source .....                                         | 131 |
| Partial Dependence Analysis .....                                             | 131 |
| E3. Private.....                                                              | 137 |
| E3.1Average Water Charges .....                                               | 137 |
| Feature Importance .....                                                      | 139 |
| E3.2 10-year Health Violations.....                                           | 146 |
| Feature Importance .....                                                      | 147 |
| Interaction Strengths (Friedman H-Statistic) .....                            | 150 |
| Partial Dependence Analysis .....                                             | 150 |
| E3.2.1 Synthesis: Conceptual Model of Private Utility Violation Drivers ..... | 152 |
| Appendix F Dashboard for data and model exploration .....                     | 154 |

# 1. Introduction

## 1.1 Motivation and Sponsorship

Access to safe and affordable drinking water is a fundamental public health necessity and a key dimension of community resilience. The United States operates one of the world's most complex drinking water governance systems: over 50,000 community water systems (CWSs) spanning public, private, and mixed ownership types serve populations ranging from fewer than 25 persons to several million. Despite the Safe Drinking Water Act's substantial achievements since 1974, persistent challenges remain. In 2015 approximately 21 million Americans were served by CWSs in active violation of health-based quality standards (Allaire et al., 2018). Water charges rose by nearly 20% in real terms between 2017 and 2023 (Teodoro and Thiele, 2024), with rate structures becoming more regressive. In this study, using 2022 municipal finance survey data, we find that the median of the ratio of annual total operating expenses to annual revenue was 0.92 for private companies and 0.91 for public utilities, but there was considerable variation in these ratios.

This study was commissioned by the Mississippi River Cities and Towns Initiative (MRCTI) to provide a national baseline for understanding water utility performance and to situate the corridor's utilities within that national context. A companion sub-report addresses the ten corridor states specifically. The national analysis presented here is designed to serve both the utility industry and the research community.

### Mississippi River Corridor — MRCTI Sub-Report

The Mississippi River corridor includes ten states: Louisiana, Minnesota, Wisconsin, Iowa, Illinois, Missouri, Kentucky, Arkansas, Tennessee, and Mississippi. Detailed state-by-state findings are in the companion MRCTI sub-report; key corridor observations appear as sidebars throughout this report.

## 1.2 Research Questions

The study addresses four central questions:

- **How do water charges, health violations, and financial performance vary** across ownership type, water source, system scale, state context, and community demographics — individually and in combination?
- **What do public and private utilities do differently**, and where do they behave similarly? Are ownership differences in performance primarily intrinsic or structural — driven by the kinds of communities and environments each type serves?
- **What is the incremental value of nonlinear machine learning methods** relative to linear regression? Where do the two methods agree, and where do nonlinear discoveries change policy-relevant conclusions?
- **Where do Mississippi River corridor utilities sit** within the national distribution on each performance dimension?

## 1.3 A Methodological Note: Survey-Based Charges

A distinguishing feature of this study is how water charges are measured. The majority of prior literature estimates households' payments by applying a utility's published rate schedule to a standard household that consumes a fixed volume of water. While this enables direct rate comparisons, it overlooks the actual distribution of bills paid by households of different types within a utility's service area.

This study instead computes average water charges and charge variability — the spread of actual charges paid, measured as the **standard deviation of annual charges** within each utility's service area — directly from the 2022 American Community Survey (ACS). The ACS collects self-reported annual household-level water expenditures; these reports are aggregated to the PWSID level using EPIC's CWS service area boundary mapping. This approach captures the *realized* heterogeneity of water costs as experienced by households, including the effects of tiered pricing, income-based assistance, multi-unit versus single-family service, and consumption diversity. Charge variability is therefore a novel outcome variable: it measures how equitably costs are distributed within a utility's customer base, not just the average level.

Throughout this report, '**charge variability**' refers to the standard deviation of annual water charges within a utility's service area — a measure of the internal spread of what customers actually pay. This is distinct from inter-utility variation in average charges.

## 1.4 Report Organization

Section 2 provides the literature review. Section 3 describes datasets, variable construction, and analytical methods. Section 4 presents exploratory analyses. Section 5 reports statistical inference results from GLMs and RF models. Section 6 synthesizes findings across methods and datasets. Appendices provide database development details (Appendix A), PCA methodology and component interpretations (Appendix B), a reference on technical statistical methods (Appendix C), details of the GLM fits (Appendix D), details of the RF Model fits (Appendix E), and access to and illustrations of the Dashboard (Appendix F).

## 2. Performance of U.S. Water Utilities: Health Violations, Charges, and Finances — A Decade of Evidence

*Contextualizing Our Findings With Respect To The Literature with an Emphasis on Ownership, Demographic Drivers, and Geographic Factors*

The decade spanning 2015–2025 produced a substantial and increasingly convergent body of empirical research on the performance of United States community water systems (CWSs) with respect to Safe Drinking Water Act (SDWA) compliance, average charges, and utility finances. This review synthesizes **eleven major studies** that draw on national- and state-level datasets, panel econometric models, machine learning methods, and geospatial analyses.

Four cross-cutting themes emerge:

First, *system scale* is the single most consistent predictor of health-based violations, with very small systems violating at rates 13 or more times those of large systems, a finding robust across all methods and geographies.

Second, the *ownership question* — whether public or private utilities perform better — remains empirically contested: publicly owned systems show more health-based violations on average, but privately owned systems exhibit more monitoring and reporting violations, and ownership effects are substantially moderated by system size, source type, and state regulatory context.

Third, *demographic disparities* in both violations and affordability are pervasive, with communities characterized by higher concentrations of racial minorities, low-income households, rural populations, and limited-English speakers disproportionately burdened across multiple outcomes and geographic scales.

Fourth, *water charges and rate structures* are rising rapidly and becoming more regressive, with fixed charges displacing volumetric rates in ways that systematically disadvantage low-consumption, low-income households.

A critical and consistent gap in the literature is that the treatment of these outcomes is treated as separable, whereas our analyses suggest that some communities face a compound disadvantage: higher charges relative to what cost-recovery logic would predict and more violations. This is a structural pattern with significant implications for regulatory design and federal infrastructure investment policy. ***In this chapter, we provide a backdrop for our findings from these 11 major studies, specifically to highlight where they agree with and contrast with our main findings.*** The differences in our study arise from newer data and analytical methods. Notably, none of the past studies have had access to utility financial data at the scale we developed.

### 2.1 Introduction

Access to safe and affordable drinking water is a fundamental public health imperative. The Safe Drinking Water Act (SDWA), enacted in 1974 and substantially amended in 1996, establishes the primary federal framework for regulating the quality of drinking water supplied by public water systems (PWSs) in the United States. Despite the SDWA's significant achievements over five decades, which extended health-based protections to more than 90 contaminants and improved average water quality for hundreds of millions of Americans (U.S. EPA, 2024), persistent challenges remain. As of 2015, approximately 21 million people were served by CWSs in violation of health-based quality standards (Allaire et al., 2018), and a growing body of evidence documents systematic disparities in compliance and affordability across demographic and geographic dimensions.

The decade from 2015 to 2025 witnessed a substantial expansion and methodological diversification of this literature, driven in part by the catalytic effect of the Flint, Michigan, lead crisis (2014–2019), which refocused national attention on the equity dimensions of drinking water governance. Simultaneously, the availability of nationally accessible administrative databases, including the EPA's Safe Drinking Water Information System (SDWIS), U.S. Census Bureau American Community Survey (ACS) data, and utility financial reporting systems, enabled more rigorous empirical inquiry than was previously possible. This period also saw the emergence of machine learning approaches (e.g., random forest models, gradient boosting) as complements to traditional regression frameworks, permitting the detection of nonlinear relationships, threshold effects, and higher-order interactions that standard linear models systematically obscure.

This review synthesizes the major findings of this decade-long research effort with an emphasis on *three dimensions*: (1) the role of *public versus private utility* ownership in shaping compliance and financial outcomes; (2) *demographic drivers* of disparate performance, particularly race, income, and urban/rural status; and (3) *geographic and state-level factors* that create systematic regional variation in utility performance beyond what observable system characteristics explain. The review is organized around four substantive sections corresponding to these themes, followed by an integrative discussion of cross-cutting findings, methodological limitations, and directions for future research. A comparative table summarizing the major studies reviewed, their data sources, sample characteristics, and analytical methods, is provided in Table 1.

## 2.2 National Trends and Geographic Patterns in Health-Based Violations

### 2.2.1 Long-Run National Patterns

The most comprehensive national assessment of health-based drinking water violations was published by Allaire et al. (2018) in *Proceedings of the National Academy of Sciences*. Analyzing a balanced panel of 17,900 CWSs drawn from the EPA's SDWIS over 34 years (1982–2015), the authors documented that between 9 and 45 million people were potentially affected by health-based violations annually over the study period, with the wide interval reflecting variation in population-weighted exposure. While total violations more than doubled over the study period, partly attributable to the addition of new contaminant regulations, the fraction of systems in active violation in any given year remained relatively stable at 3–10%. Spatial clustering of violations was pronounced, with *persistent hotspots* identified in *Texas, Oklahoma, and Idaho*, a geographic pattern that has since been confirmed and extended by subsequent analyses employing different methodological frameworks.

Allaire et al. (2018) identified *system scale, source water type (groundwater versus surface water), and socioeconomic context as the primary correlates of violation risk*. Rural systems serving low-income communities of color were found to have substantially higher rates of total coliform violations. The finding that *violations were concentrated in smaller, groundwater-dependent, economically disadvantaged systems* has been replicated across every subsequent major national study reviewed here, constituting perhaps the most robust empirical regularity in the U.S. drinking water literature.

An important methodological contribution of Allaire et al. (2018) was the explicit separation of health-based violations, exceedances of maximum contaminant levels (MCLs), treatment technique violations, and disinfectant residual violations from monitoring and reporting (M&R) violations. The latter are recognized as having poor data quality in the national SDWIS database, with only 9–23% of M&R violations estimated to be reported to the national system (Allaire et al., 2018), resulting in a systematic underestimation of compliance challenges in small systems that are most likely to underreport. This distinction is critical for cross-study comparisons and for regulatory interpretation of apparent ownership effects (see Section 3). In

the current work, we analyzed both health and total violations but focused primarily on health violations, given the reporting concerns noted in the literature.

## 2.2.2 Geographic Concentration and State-Level Effects

The geographic concentration of violations is not merely a reflection of demographic and system characteristics: substantial state-level residual variation persists even after controlling for all observable covariates. Fu et al. (2020) explicitly demonstrated that municipal-level heterogeneities, reflecting local regulatory cultures, enforcement capacity, and economic conditions, explain a substantial fraction of the variance in CWS violation behavior across the United States. *This finding aligns with evidence from our random forest modeling of public utility violations across three national datasets (EMMA, public utilities, private utilities), in which state identity variables (particularly Oklahoma and Louisiana for health-based violations; California and Washington for average charges) retained large, statistically robust effects even after conditioning on rich sets of demographic, financial, and source-type predictors.*

Oklahoma's status as a persistent violation outlier, confirmed across multiple independent analyses and methodological frameworks (Allaire et al., 2018; Fu et al., 2020; this report) is not reducible to its groundwater dependence or demographic profile alone. The state's violation excess in public utilities persists after controlling for these covariates, implying a true state-specific regulatory and geological environment: a combination of oil-and-gas industry contamination, agricultural runoff, and the density of small, undercapitalized public systems. ***Critically, this Oklahoma effect is concentrated in the public utility sector; private utilities operating in Oklahoma show no statistically distinguishable violation excess in financial-structure-controlled models (this report), suggesting that the compliance deficit is an organizational and institutional property of the public sector in this state rather than a geologically universal exposure.***

The role of water source type as both a direct predictor and a dominant interaction variable deserves particular emphasis. Allaire et al. (2018) documented that groundwater-dependent systems face higher violation rates than surface water systems in the national mixed-ownership sample, consistent with the particular contamination vulnerabilities (naturally occurring arsenic, nitrates, petroleum-related compounds) of small private wells and community groundwater systems in the South-Central states. ***However, this relationship reverses in the national public utility subsample, where surface water systems average 3.4 times more violations than groundwater systems (present analysis)***. This context dependence of the source-type effect, one of the clearest examples of a nonlinear, interaction-conditioned relationship, illustrates the interpretive risks of applying national-average coefficients to specific utility subpopulations.

## 2.3 Public versus Private Ownership: A Contested Empirical Question

### 2.3.1. The State of the Evidence

The question of whether public or private ownership produces better drinking water outcomes has attracted persistent scholarly attention since at least the foundational work of Wallsten and Kosec (2008), who analyzed a national panel from 1997–2003 and found *minimal average ownership effects after controlling for source water, location, income, and urbanization*. The decade reviewed here has not resolved the question but has substantially refined its framing, identifying conditions under which ownership matters and mechanisms through which it operates.

The most recent comprehensive study specifically focused on ownership is Fu, Liu, and Swallow (2020), who constructed a comprehensive national violation dataset from SDWIS covering 2011–2017. Using negative binomial panel regression with municipal-level fixed effects, *they found that publicly owned PWSs*

*commit significantly more MCL, treatment technique, and health-related violations than privately owned PWSs, while private systems incur more monitoring and reporting violations.* This directional pattern, public systems worse on health-based metrics, private systems worse on administrative metrics, is consistent across multiple studies and methods (Switzer & Teodoro, 2017, 2018; Konisky & Teodoro, 2016) and has been interpreted as reflecting different organizational incentive structures: *publicly managed utilities operating under democratic accountability may be more responsive to constituent pressure to avoid administrative penalties (M&R violations), while investor-owned utilities operating under rate-of-return regulation have stronger financial incentives to minimize treatment investment costs, potentially at the expense of MCL compliance.*

Importantly, Fu et al. (2020) emphasized that the *ownership effect is substantially attenuated* and, in some specifications, reversed once municipal-level heterogeneities are controlled for. This finding implies that much of what appears as an ownership effect in simpler models is, in fact, a geographic and demographic sorting effect: publicly owned utilities are more likely to serve rural, small, low-income communities (the very characteristics associated with more violations regardless of ownership), and naive ownership comparisons confound these compositional differences with true causal effects of governance structure. ***This interpretation is consistent with the random forest models developed in the current report, in which ownership type explains a very small fraction of total variance in both charges and violations once demographic, financial, source-type, and geographic predictors are included.***

The EMMA GLM analysis in the current study confirms the direction of the finding by Fu et al. (2020): ***public utilities charge approximately \$41 more per year than private utilities after controlling for all covariates. However, the ownership effect on violations is negligible once source type, state, scale, and demographics are controlled,*** consistent with the broader literature's conclusion that ownership effects are secondary to structural factors. This is not a surprise since the GLM, while applied to a different data set, is also linear. However, ***the nonlinear RF analysis adds a new dimension: public and private utilities differ not just in average levels but in the functional form of the relationship between predictors and outcomes.***

### 2.3.2. Ownership and Charges: A Different Story

The relationship between ownership and water charges is less extensively studied than the violations literature, but it has received renewed attention as concerns about water affordability have mounted. *The EMMA (mixed-ownership) generalized linear model in our analysis finds that publicly owned utilities charge approximately \$41 more per year than investor-owned utilities of comparable characteristics.* This finding is consistent with the broader literature showing that public utilities, which must finance infrastructure through municipal debt rather than equity, face structurally different cost-recovery pressures. Our random forest analysis adds important nuance: ***this \$41 average masks a distribution ranging from approximately –\$30 (Mississippi and Louisiana public utilities charging well below the national mean) to +\$50 (California, Washington, and Oregon public utilities charging well above), with state identity explaining approximately 20% of total charge variance, more than any other predictor group.***

*Private utility water charges show rather different trends.* Our random forest analysis of private utility charges finds that charges are effectively regime-based rather than proportional: nearly all predictors exhibit step-function relationships for average charges, with demographic market segmentation, specifically the concentration of knowledge-economy workers, economically affluent residents, and urban diversity, defining the tier assignment. Operating costs per customer do not interact significantly with these demographic premiums in 2D partial dependence plots, indicating that private utility charges are driven by market position, financial strategy, and state regulatory environments. *Pennsylvania and California create*

*irreducible charge premiums in private utility models, consistent with the large and heavily regulated investor-owned utility sectors in both states.*

*A critical finding in our work reported here, which bridges the ownership and affordability literature, is that the PCF2 (Tax Impact) financial principal component is the leading predictor of both private utility charges and private utility violations, with the same positive sign in both models. Tax-dominant investor-owned utilities charge more and violate more, a compound burden on ratepayers that cannot be attributed to either pricing pressure or compliance failure in isolation. Whether PCF2 reflects genuine strategic differences or serves as a proxy for a multi-system corporate holding company structure requires further investigation using matched organizational data that was not available.*

## 2.4. Demographic Drivers of Disparate Outcomes

### 2.4.1. Race, Ethnicity, and Health-Based Violations

The relationship between community racial and ethnic composition and SDWA compliance is one of the most extensively replicated findings in the literature reviewed here. Switzer and Teodoro (2017) used national utility data to demonstrate that *SDWA violations are greater in low-income communities with higher proportions of Black and Hispanic residents, identifying race/ethnicity as a stronger sociodemographic predictor than income in multivariate models*. Their follow-up study (Switzer & Teodoro, 2018) advanced the analysis by modeling the interactive effects of race, ethnicity, and socioeconomic status, finding that the racial composition of a community is a predictor of drinking water quality even after controlling for income, and that violations in low-SES communities are most strongly predicted by the percentage of Black and Hispanic residents. McDonald and Jones (2018) replicated and extended these findings at the county level using national SDWIS data from 2011–2015, finding that counties with higher proportions of underrepresented minorities and uninsured households faced elevated odds of both initial and repeat health-based violations, and calling for greater attention to the equity dimensions of federal SDWA enforcement targeting.

Allaire and Acquah (2022) provided the most methodologically rigorous *state-level replication of these findings for California, using areal weighting to assign service-area demographics more precisely than the county-level proxy approach used in earlier studies*. Their probit panel analysis of 1,693 California CWSs from 2000–2018 found violation rates several times higher at systems serving low-income areas and communities of color, and demonstrated that these disparities persist even in models stratified by system size, which is a critical finding because system size is the strongest predictor of violations and is positively correlated with demographic disadvantage. The persistence of racial and income disparities across size classes implies that demographic context carries independent information about compliance risk beyond what system scale alone can explain.

Marcillo et al. (2021) extended this analysis to Virginia, applying a novel GIS-based service-area delineation at the ZIP code scale to link 2000 Census demographics with 2006–2016 SDWA violations. Using negative binomial regression, they found a significant positive association between the proportion of Black residents in a CWS service area and the likelihood of health-based violations, even after controlling for rurality and system characteristics. Importantly, Marcillo et al. (2021) also documented that the spatial analysis method selected for environmental justice studies can substantially affect results and conclusions — a methodological caution with implications for how national analyses that rely on county-level census proxies should be interpreted.

### 2.4.2. Income, Poverty, and the Affordability Paradox

The relationship between community income levels and water system outcomes is more complex than a simple directional association would suggest. *For compliance, the preponderance of evidence confirms that*

*lower-income communities face more violations, consistent with infrastructure underinvestment, limited technical capacity, and reduced political advocacy that characterize economically distressed utility service areas. **Our random forest models of public utility violations confirm this relationship as a monotonically decreasing function of Economic Affluence, with the steepest decline occurring in the poverty tail, and a flat plateau at above-average income levels, suggesting that economic resources above the poverty threshold are sufficient to eliminate the income-driven compliance deficit.***

*For charges, the relationship is more counterintuitive. Our public utility random forest charges model finds that the *Economic Affluence predicts higher average charges, not lower, with a continuous, accelerating positive effect that constitutes the largest marginal demographic range in the public utility charges analysis (approximately \$60 from most to least poor service areas). Both the GLM (positive coefficient on poverty-related principal components) and the RF (monotone increasing partial dependence) confirm this positive poverty–charges association in public utilities, consistent with the infrastructure economics interpretation: aging, dispersed systems serving persistent-poverty communities cost more per connection to maintain, and lower per-connection consumption reduces revenue without reducing fixed costs, forcing utilities to charge more to recover costs. This finding, that public utilities serving the poorest communities charge more after adjusting for other factors, AND experience more violations, constitutes a compound disadvantage that is not captured by either the affordability or the compliance literature in isolation.**

The national affordability literature has documented rising average. Teodoro and Thiele (2024) analyzed biennial rate data from approximately 6,200 utilities from 2017 to 2023, finding that average combined monthly water and sewer bills rose from \$79.39 to \$95.02 (a 19.7% increase), and that rate structures became more regressive over this period, *with fixed charges increasing approximately 45% while volumetric charges declined 8%.* These structural shifts, which the authors link to utilities' incentives to reduce revenue risk by collecting more through fixed charges, *disproportionately burden low-volume, low-income households for whom fixed charges represent a larger share of total bills.* The EPA's (2024) Water Affordability Needs Assessment estimated that between 12.1 and 19.2 million U.S. households lack affordable access to water services, with *financial burden concentrated in small, rural, and Southern utilities, precisely the systems that also carry the highest violation rates.*

### **2.4.3. Rural Status and System Scale**

*System scale, operationalized as population served, number of service connections, or aggregate financial volume, is the single most consistent predictor of health-based violations across the decade of literature reviewed here. Dobbin et al. (2023), analyzing consolidation activity in California in *Water Resources Research*, documented that very small systems violate SDWA standards at a rate more than 13 times that of large systems, a ratio consistent with national estimates. This extreme scale-violation gradient reflects the convergence of multiple mechanisms: small systems lack economies of scale in infrastructure investment, cannot justify full-time professional operational staff, have limited financial capacity to fund treatment upgrades and monitoring equipment, and face weaker regulatory attention due to sheer numerical abundance (over 50,000 CWSs nationally, the vast majority classified as "small" by EPA).*

Importantly, scale-protection effects are ownership- and context-dependent in ways that challenge universal prescriptions. For public utilities, both the full-utility panel models (Allaire et al., 2018; present analysis) and our machine learning analyses confirm that larger systems have significantly fewer violations, consistent with the standard scale-economies interpretation. For private utilities, our findings were similar. However, an initial analysis at the company level indicated that some of the larger private utilities had a high number of violations. Upon investigation, we discovered that, at the PWSID level, those violations typically came from the smaller, poorer utilities owned by the company, and that the company-level result

was a consequence of aggregation. Thus, for violations, there are no significant differences by ownership with respect to rural status and system scale, and we retain the comprehensive findings.

## 2.5. Financial Determinants and Infrastructure Investment

### 2.5.1. Debt, Revenue Recovery, and Public Utility Charges

The financial structure of public utilities is a substantially underexplored determinant of both charge levels and compliance outcomes, given its policy importance. *Our public utility random forest charges model analyzed in the present study identifies debt leverage (PCF7 — total debt outstanding relative to revenue generation) as the largest single financial predictor of charges, with a step-function relationship of approximately \$46 per year at the debt-to-revenue threshold: utilities carrying heavy debt relative to their revenue base must set rates substantially higher to service their obligations. This finding is consistent with the debt servicing interpretation of public utility rate-setting but adds important functional form information: the relationship is not proportional (a linear premium per dollar of debt) but binary (a threshold effect), implying that the policy-relevant question is identifying which utilities have crossed a debt-revenue threshold rather than targeting the marginal debt level.*

Hansen and Mullin (2022), in a survey-based analysis of U.S. local elected officials, documented that financial barriers are the primary stated obstacle to water infrastructure investment: officials weigh capital project costs against debt burden and rate affordability concerns, with fear of political opposition to rate increases functioning as a binding constraint. *This behavioral finding is consistent with the econometric pattern identified in our random forest models: public utilities facing high debt burdens compensate through higher charges, while those with revenue shortfalls (negative PCF5, Revenue Expense Contrast) also raise rates to close the gap — a step-function relationship at zero Revenue Expense Contrast.* Both mechanisms are financially rational *but compound the burden on ratepayers in communities with limited capacity to absorb rate increases.*

***A non-monotone financial relationship identified in our random forest analysis of public utility violations — where violations are highest at both financial deficit and surplus extremes, with minimum violations at the revenue-expense breakeven point — represents a finding with no clear parallel in the existing literature.*** The standard assumption that financially stronger utilities comply better would predict a monotone negative relationship between revenue surplus and violations. The RF's identification of a U-shaped response suggests that revenue extraction above breakeven may be associated with under-investment in operational maintenance and compliance infrastructure, a finding that warrants replication with panel data before informing regulatory policy but which challenges the use of financial health metrics (revenue surplus, profit margin) as compliance proxies in regulatory oversight frameworks.

### 2.5.2. Consolidation as a Policy Response

The literature increasingly converges on water system consolidation — the formal merger of two or more previously independent systems — as the structural policy response most likely to address the scale-compliance nexus for public utilities. Dobbin et al. (2023), in the most systematic empirical analysis of consolidation to date, analyzed over 200 consolidation projects completed in California between 2015 and 2021, finding that consolidations preferentially target the highest-risk, most underperforming small systems. Water quality compliance concerns were cited as the primary motivation in 81% of projects with documented rationale; climate change impacts (drought, fire, landslide) were the second-most-common motivation (>10%). Critically, Dobbin et al. (2023) also found that receiving systems — the entities absorbing consolidated systems — are themselves commonly small (45% serving fewer than 5,000

customers), challenging the assumption that consolidation necessarily delivers large-scale operational efficiency gains.

Based on an empirical analysis of SDWA compliance data, utility performance indicators, and operational efficiency metrics across the national CWS population, Teodoro (2024, in the Springer *Safe Drinking Water Act at Fifty* volume) argues that consolidating the American drinking water sector to approximately 6,000 utilities with a minimum size of 20,000 connections would deliver better health outcomes, greater efficiency, more effective regulation, and lower prices than the current fragmented landscape of over 50,000 systems. This analysis provides a scale threshold estimate — 20,000 connections — consistent with, though somewhat larger than, the regulatory and economic literature on minimum efficient scale in water utilities, and establishes a quantitative benchmark for evaluating consolidation targets. Our analysis does not shed additional light on the literature.

## 2.6. Synthesis and Discussion

### 2.6.1. What the Literature Establishes with Confidence

The body of evidence reviewed here supports **five findings with high confidence** across multiple independent studies, datasets, and analytical methods:

**First**, *system scale is the dominant predictor of health-based violations in the public utility sector, with the scale-violation relationship exhibiting a strongly nonlinear (accelerating) form that gives very small systems a disproportionate share of the compliance burden.*

**Second**, *communities characterized by higher proportions of racial minorities, low-income households, and rural residents face consistently and persistently higher rates of health-based violations — a pattern that persists across geographic scales, time periods, and analytical methods, though its magnitude varies with the precision of service area demographic assignment.*

**Third**, *water charges have risen substantially in real terms over the past decade, with rate structures becoming more regressive as volumetric charges are displaced by fixed charges, worsening affordability, particularly for low-volume, low-income households.*

**Fourth**, *state-level regulatory and environmental context creates residual variation in both violations and charges that is not explained by observable utility or community characteristics, implying that state-level governance reforms are necessary (though not sufficient) to address national compliance disparities.*

**Fifth**, *water source type is a major moderator of violation risk, with the direction of source-type effects depending critically on the composition of the utility population under study.*

### 2.6.2. Areas of Persistent Uncertainty

Despite this convergence on broad patterns, several important questions remain contested or underexplored. The causal direction of the poverty–charges relationship in public utilities — whether high charges reflect higher costs in poor communities (infrastructure economics) or reflect other political-economic dynamics — cannot be established from cross-sectional or static panel data and requires natural experiments or longitudinal rate-change analysis. The ownership effect on health-based violations remains sensitive to control variables, the analytical unit, and the time period, and the mechanisms by which ownership influences compliance (incentive structures, professional capacity, political accountability) have not been empirically decomposed with sufficient rigor. The nonlinear relationships identified through our machine learning approaches — *threshold effects in debt leverage and revenue expense contrast, regime-switching in private utility pricing, multiplicative source-type × poverty interactions in public utility violations*

— represent genuine discoveries that challenge linear model assumptions, but most have not been validated on alternative variable constructions.

### 2.6.3. A Compound Disadvantage Framework

The most important conceptual contribution of integrating the violations, charges, and financial performance literatures is the identification of a *compound disadvantage* structure that *has not been explicitly articulated in the existing literature*. The communities and utility service areas that face the most health-based violations — small, rural, groundwater-dependent, and demographically disadvantaged — also tend to pay higher charges in the public utility sector (per the positive poverty–charge gradient documented here) and receive less responsive enforcement of SDWA violations (per Pullen Fedinick et al., 2021). This compound structure ***means that violations and affordability are not independent problems amenable to independent solutions: they are co-determined by the same infrastructure economics, demographic context, and governance dynamics. Policy frameworks that address affordability through rate suppression without investing in compliance will not break the cycle of compounding disadvantage; frameworks that address compliance without affordability assistance will impose unaffordable burdens on the communities most in need of infrastructure improvements.***

## 2.7 Conclusions and Research Priorities

Ten years of increasingly rigorous empirical research have substantially clarified, but not fully resolved, the determinants of U.S. water utility performance with respect to health-based compliance and charges. Several research priorities emerge from this review.

**First**, the causal mechanisms linking demographic context to utility performance require identification strategies beyond the cross-sectional and observational designs that dominate the current literature. Regression discontinuity designs exploiting regulatory thresholds (e.g., EPA system size categories), difference-in-differences analyses of consolidations or ownership transfers, and synthetic control methods applied to state regulatory reforms would advance causal inference substantially if data to support such analyses were available. The strong “state” effects across all our analyses suggest that additional national information is needed to generalize the findings beyond data from any single state.

**Second**, the convergent finding that public utility charges are positively associated with service area poverty levels warrants dedicated investigation using panel data to isolate the mechanisms (infrastructure age, connection density, consumption patterns, political economy of rate-setting) through which this counterintuitive relationship operates.

**Third**, the nonlinear and interaction-conditioned patterns identified through machine learning approaches (threshold effects, regime switching, multiplicative source-type × poverty interactions) constitute functional-form discoveries of policy relevance that have not been incorporated into standard regulatory assessment frameworks. Replication of these nonlinear findings using alternative machine learning architectures and sensitivity analyses on variable construction is a priority for the field.

**Fourth**, the compound disadvantage outcome suggests that *regulatory frameworks linking rate adequacy, infrastructure investment, and compliance outcomes are likely to be more effective than the current administrative separation of affordability (state public utility commission jurisdiction) from compliance (EPA/state primacy agency jurisdiction)*. Empirical evaluation of states or utilities that have adopted more integrated approaches would provide evidence for this policy architecture.

**Finally**, the literature's consistent finding that very small public water systems are the single highest-risk archetype, for violations, financial distress, and affordability burden, combined with the Bipartisan

Infrastructure Law's substantial new funding for drinking water, suggests that tracking the compliance and affordability outcomes of Bipartisan Infrastructure Law-funded consolidations and infrastructure investments over the coming decade offers an important natural experiment for evaluating the causal effects of scale and investment on the outcomes documented in the studies reviewed here.

## 2.8 What This Study Adds

This study extends the literature in four ways. *First*, it uses actual survey-based charges and charge variability rather than rate-formula estimates. *Second*, it applies nonlinear machine learning alongside GLMs, enabling the detection of threshold effects, regime switching, and interaction structures that prior linear analyses miss. *Third*, it analyzes private utility financial structure at a granularity not previously available in academic research, revealing that private utility pricing and violations are organized by financial structure in ways fundamentally different from public utilities. *Fourth*, it provides the first comprehensive national analysis of the effects of public utility financial structure on both charges and violations, using per-capita-scaled municipal finance principal components.

## 2.9 References

- Allaire, M., Acquah, S. (2022). Disparities in drinking water compliance: Implications for incorporating equity into regulatory practices. *AWWA Water Science*, 4(3), e1274. <https://doi.org/10.1002/aws2.1274>
- 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>
- Dobbin, K., McBride, J., & Pierce, G. (2023). Panacea or placebo? The diverse pathways and implications of drinking water system consolidation. *Water Resources Research*, 59, e2023WR035179. <https://doi.org/10.1029/2023WR035179>
- 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>
- Hansen, K., & Mullin, M. (2022). Barriers to water infrastructure investment: Findings from a survey of U.S. local elected officials. *PLOS Water*, 1(8), e0000039. <https://doi.org/10.1371/journal.pwat.0000039>
- Marcillo, C., Krometis, L.-A., & Krometis, J. (2021). Approximating community water system service areas to explore the demographics of SDWA compliance in Virginia. *International Journal of Environmental Research and Public Health*, 18(24), 13254. <https://doi.org/10.3390/ijerph182413254>
- 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>
- Pullen Fedinick, K., Taylor, S., Roberts, M., & Olson, E. (2021). *Watered Down Justice*. Natural Resources Defense Council, Environmental Justice Health Alliance for Chemical Policy Reform, and Coming Clean. <https://www.nrdc.org/resources/watered-down-justice>
- Switzer, D., & Teodoro, M. P. (2017). The color of drinking water: Class, race, ethnicity, and Safe Drinking Water Act compliance. *Journal – American Water Works Association*, 109(9), 40–45. <https://doi.org/10.5942/jawwa.2017.109.0128>
- Switzer, D., & Teodoro, M. P. (2018). Class, race, ethnicity, and justice in safe drinking water compliance. *Social Science Quarterly*, 99(2), 524–535. <https://doi.org/10.1111/ssqu.12397>

- Teodoro, M. P. (2024). The scale imperative. In M. P. Teodoro (Ed.), *The Safe Drinking Water Act at Fifty*. Springer Nature. [https://doi.org/10.1007/978-3-032-10830-2\\_6](https://doi.org/10.1007/978-3-032-10830-2_6)
- Teodoro, M. P., & Thiele, R. (2024). Water and sewer price and affordability trends in the United States, 2017–2023. *Journal – American Water Works Association*, 116, 14–24. <https://doi.org/10.1002/awwa.2315>
- U.S. Environmental Protection Agency. (2024). *Water Affordability Needs Assessment: Report to Congress*. Office of Water. <https://www.epa.gov/waterfinancecenter/water-affordability-needs-assessment>
- Wallsten, S., & Kosec, K. (2008). The effects of ownership and benchmark competition: An empirical analysis of U.S. water systems. *International Journal of Industrial Organization*, 26(1), 186–205. <https://doi.org/10.1016/j.ijindorg.2006.10.002>

**Table 2.1 Summary of Major Studies: Data Sources, Sample Characteristics, and Analytical Methods**

| Study                                             | Data source                                                 | Sample                                      | Period    | Methods                                                                                            | Key findings (ownership / demographics / geography)                                                                                                                                                                                                             |
|---------------------------------------------------|-------------------------------------------------------------|---------------------------------------------|-----------|----------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Allaire, Wu & Lall (2018) PNAS                    | EPA SDWIS; U.S. Census county data                          | 17,900 CWSs, balanced panel                 | 1982–2015 | Logistic & OLS panel regression; spatial clustering; county-level fixed effects                    | 9–45 M people affected annually. Violations hot-spotted in TX, OK, ID. Low-income, rural, communities of color face higher rates. Groundwater systems and small systems most at risk.                                                                           |
| Switzer & Teodoro (2017) J. AWWA                  | EPA SDWIS; ACS demographics                                 | National utility data, all CWSs >1,000 pop. | 2010–2015 | Negative binomial regression; state fixed effects                                                  | Violations greater in low-income communities with higher Black and Hispanic populations. Race/ethnicity strongest sociodemographic predictor. Public utilities have more health-based violations than private.                                                  |
| Switzer & Teodoro (2018) Social Science Quarterly | EPA SDWIS; ACS; local government data                       | U.S. local government CWSs >1,000 pop.      | 2010–2013 | Negative binomial; race × SES interaction models                                                   | Interactive effects of race/ethnicity and SES. Violations in low-SES areas strongly predicted by % Black and Hispanic. Public ownership associated with more health-based violations; interaction with political representation.                                |
| McDonald & Jones (2018) Am. J. Public Health      | EPA SDWIS; ACS; county health data                          | National county-level analysis, all CWSs    | 2011–2015 | Multivariable logistic regression; odds ratios for initial and repeat violations                   | Communities of color and low-income counties face greater odds of health-based violations. Repeat violations more prevalent in minority and uninsured counties. Small systems disproportionately affected regardless of race.                                   |
| Fu, Liu & Swallow (2020) Agr. & Res. Econ. Review | EPA SDWIS; ACS; county economic data                        | Comprehensive national dataset of CWSs      | 2011–2017 | Negative binomial panel regression; municipal-level fixed effects                                  | Publicly owned CWSs commit significantly more MCL, treatment technique, and health-based violations. Private systems have more monitoring/reporting violations. Municipal heterogeneity (local regulations, economic conditions) explains substantial variance. |
| Allaire & Acquah (2022) AWWA Water Science        | EPA SDWIS; ACS areal-weighted demographics; CA primacy data | 1,693 CWSs in California                    | 2000–2018 | Probit panel regression; areal-weighting for service area demographics; size stratification        | Violation rates several times higher for systems serving low-income areas and communities of color, even when stratified by size. Size does not eliminate equity disparities. Regulatory assessment underestimates inequity by ignoring demographic context.    |
| Teodoro & Thiele (2024) J. AWWA                   | AWWA biennial rate survey; Census income data               | ~6,200 utilities nationally                 | 2017–2023 | Descriptive panel analysis; affordability threshold (4.5% MHI); fixed vs. volumetric decomposition | Average combined monthly bill rose from \$79 to \$95 (2017–2023). Rate structures became more regressive: fixed charges +45%, volumetric -8%. Low-income affordability worsened, driven by extreme unaffordability at a minority of utilities.                  |

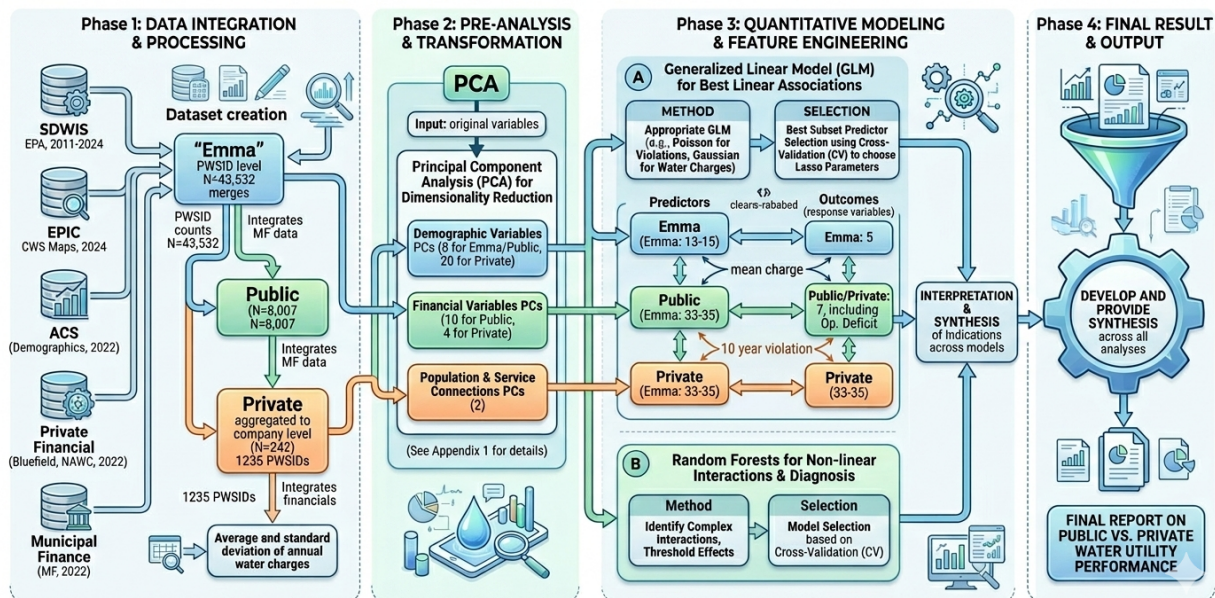
| Study                                                             | Data source                                         | Sample                                       | Period    | Methods                                                                                            | Key findings (ownership / demographics / geography)                                                                                                                                                                                                         |
|-------------------------------------------------------------------|-----------------------------------------------------|----------------------------------------------|-----------|----------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Dobbin et al. (2023)<br>Water Resources Research                  | CA State Water Board consolidation records; SDWIS   | 200+ consolidation projects in CA, 2015–2021 | 2015–2021 | Descriptive analysis; discrete choice modeling of consolidation targets                            | Small systems violate SDWA at 13× rate of large systems. Consolidations preferentially target highest-risk small systems. Climate change (drought, fire) is second-most-cited consolidation driver. Policy reforms critical to advancing access and equity. |
| Pullen Fedinick et al. (2021)<br>NRDC/EJHA (Watered Down Justice) | EPA SDWIS; CDC SVI; ACS                             | National county-level; all CWSs              | 2016–2019 | Correlation analysis; vulnerability index; enforcement response analysis                           | Communities of color, low-income, and limited-English-proficiency populations have higher violation rates. Race/ethnicity strongest predictor of slow/inadequate enforcement. Disparities in EPA response time by community demographics.                   |
| EPA Water Affordability Needs Assessment (2024)                   | AWWA rate surveys; ACS; Census income distributions | National; 787+ largest utilities             | 2017–2023 | Affordability threshold analysis (3% and 4.5% MHI); household income distribution modeling         | 12–19 M households lack affordable water access nationally. Financial gap estimated \$5–9 B/yr. Largest burden in small, rural, and Southern utilities. No permanent federal low-income assistance program exists.                                          |
| Marcillo et al. (2021)<br>Int. J. Environ. Res. Public Health     | SDWIS; ACS; GIS zip-code delineation                | Virginia CWSs                                | 2006–2016 | Negative binomial regression; GIS service area delineation; spatial environmental justice analysis | % Black residents positively associated with health-based violations. Rural and small systems (<200 connections) most likely to violate. Spatial method choice significantly affects environmental justice conclusions.                                     |

Notes: CWS = community water system; SDWIS = Safe Drinking Water Information System; ACS = American Community Survey; MCL = maximum contaminant level; SES = socioeconomic status; MHI = median household income; RF = random forest; GLM = generalized linear model.

### 3. Data Sources, Datasets, and Analytical Methods

By integrating multiple national datasets, harmonizing them at the utility (PWSID) level, and applying both statistical and machine-learning methods, the study examines how ownership structure, demographic context, financial conditions, and system characteristics shape customer charges, regulatory compliance, and operational performance. The workflow is designed to inform a transparent, evidence-based understanding of where public and private utilities differ, where they behave similarly, and what underlying factors drive those patterns. The overall workflow is illustrated below.

#### WORKFLOW FOR ANALYSING PERFORMANCE OF PUBLIC VS. PRIVATE WATER UTILITIES IN THE USA



**Figure 1** Workflow for the Analysis. Data set creation and Processing to Modeling and Synthesis

#### 3.1 Data Sources

Five primary data sources were integrated at the utility (PWSID) level:

- **EPA Safe Drinking Water Information System (SDWIS), 2011–2024.** Ownership type, water source code, population served, number of service connections, and counts of health and total violations for each community water system. Health violations include Maximum Contaminant Level exceedances, treatment technique violations, and disinfectant level violations. Total violations include monitoring and reporting violations.
- **EPIC Community Water System Maps (2024).** Geospatial CWS service area boundaries that enable population-area-weighted assignment of ACS block-group data to each PWSID and linkage to municipal finance boundaries.
- **2022 American Community Survey (ACS).** Demographic variables at the block-group level were aggregated to each PWSID using EPIC's service area mapping. The ACS also reports household water expenditure as a categorical distribution, enabling computation of the mean and standard deviation of annual charges paid within each utility's service area.
- **2022 Bluefield Research / NAWC Financial Reports.** Annual financial statements for investor-owned utilities, including operating revenue, operating expenses, total operating expenses, and income taxes. Reports were sourced as PDFs — many handwritten and scanned — digitized, verified, and updated to 2022 dollars using annual CPI inflation rates. Final analytical sample: 242 companies with 1,235 PWSIDs.

- **2022 Municipal Finance Data (Willamette University).** Government finance data for local governments matched to CWS boundaries: total utility revenue, water utility revenue, capital outlays, construction, interest expense, current expenses, and total debt outstanding.

### 3.2 Three-Tier Dataset Structure

The study employs a nested 'Russian Doll' structure, balancing breadth with financial depth:

- **EMMA (N=43,532).** The broadest layer, integrating SDWIS, EPIC, and ACS for all CWSs for which demographic data could be matched. Eight demographic principal components (PCD1–PCD8) and two scale components (PCS1–PCS2) serve as predictors. No financial data.
- **Public (N=8,007).** EMMA utilities matched to municipal finance data. Adds 10 aggregate financial PCs (PCF1–PCF10) and 10 per-capita scaled financial PCs (PCFS1–PCFS10), in addition to the EMMA demographic and scale components.
- **Private (N=1,235 PWSIDs / 242 companies).** EMMA utilities matched to Bluefield/NAWC financial reports. Company-level data aggregated across PWSIDs. Uses 20 demographic PCs (expanded from EMMA's 8 to enable finer market-segment analysis), 4 aggregate financial PCs (PCF1–PCF4), and 4 per-capita scaled financial PCs (PFS1–PFS4).

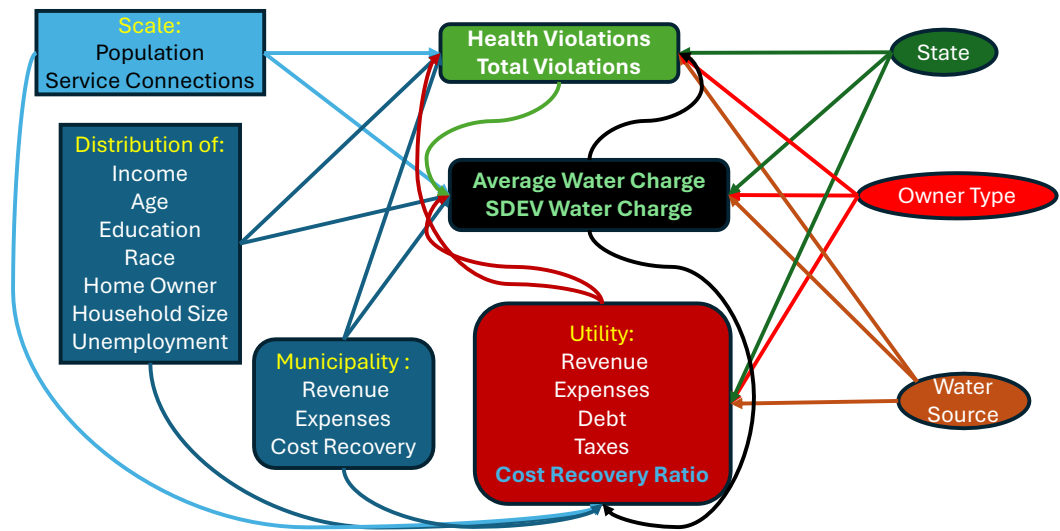
Public and private financial datasets were not combined given significant imbalance (8,007 versus 242 units) and differences in financial reporting structures between municipal and investor-owned contexts.

### 3.3 Response Variables and Modeling Structure

Seven response variables were estimated (availability varies by dataset):

- **Average annual water charge (\$/year).** Mean ACS-reported household annual water expenditure within the utility's service area.
- **Charge variability (\$/year).** Standard deviation of ACS-reported household annual water expenditure — a novel measure of within-utility distributional equity. Referred to as 'charge variability' throughout; this is the standard deviation of charges across households within a utility's service area.
- **5-year health violations (count).** Cumulative SDWA health violations 2020–2024.
- **10-year health violations (count).** Cumulative SDWA health violations 2015–2024.
- **5-year total violations (count).** Cumulative total SDWA violations 2020–2024.
- **10-year total violations (count).** Cumulative total SDWA violations 2015–2024.
- **Operating deficit indicator (binary).** Equal to 1 if annual operating expenses exceed annual revenue; 0 otherwise. Public and private datasets only.

The conceptual structure associated with the models is illustrated below. The analysis follows a logical progression from developing uncorrelated predictor groups to predictive modeling, ensuring that multicollinearity—a common headache in demographic and financial data—is handled before the analyses. Regression and Machine Learning Models are pursued with calibration in a predictive context, but we do not assume that the results are causal; rather, they describe the settings in which certain outcomes are found. To clarify, if a model predicts that higher household income is a significant predictor of health violations, we recognize that health violations in higher-income areas differ from those in lower-income areas, but we do not argue that income causes or leads to violations in any way. It is this mapping of setting and outcome that we are interested in by relating the different variables. The question being answered is where one would find x, rather than what caused x.



**Model Structure Explored for Prediction using Regression and Machine Learning Models**

**Figure 2** *Model Structures Explored*

### 3.4 Principal Component Analysis

Demographic and financial variables are strongly collinear within each group. PCA was applied separately to each dataset and variable group, yielding uncorrelated components that capture the principal axes of variation. The PC structure differs across datasets, so verbal descriptions of each component are used throughout to ensure comparability. Appendix B provides full PCA loading tables and component interpretations.

For EMMA, 8 demographic PCs span Economic Affluence (PCD1) and Urban Diversity and Wealth (PCD2) through Homeownership Strength, Deep Poverty Gradient, Demographic Homogeneity, Income Inequality, Median Income Variance, and Middle Income Domination. The public dataset uses a different 8-component structure emphasizing Economic Affluence (PCD1), Urban Density and Rental Wealth (PCD2), and vulnerability dimensions. The private dataset employs 20 demographic components, enabling finer market-segment analysis.

### 3.5 Generalized Linear Models with Lasso Regularization

GLMs with Lasso regularization and K-fold cross-validation were applied to select predictors and estimate coefficients. Gaussian link for charges, Poisson log-link for violation counts, and Logistic link for the operating deficit indicator. Predictors include state indicators (48 states; Alaska as reference), ownership type, primary source code, scale PCs, demographic PCs, and financial PCs where available, plus cross-response predictors. The Lasso penalty  $\lambda$  was selected by cross-validated mean squared error. Results are reported as regression coefficients significant at the 5% level, partial  $R^2$  by predictor group, and state coefficient maps. See Appendix C for method details.

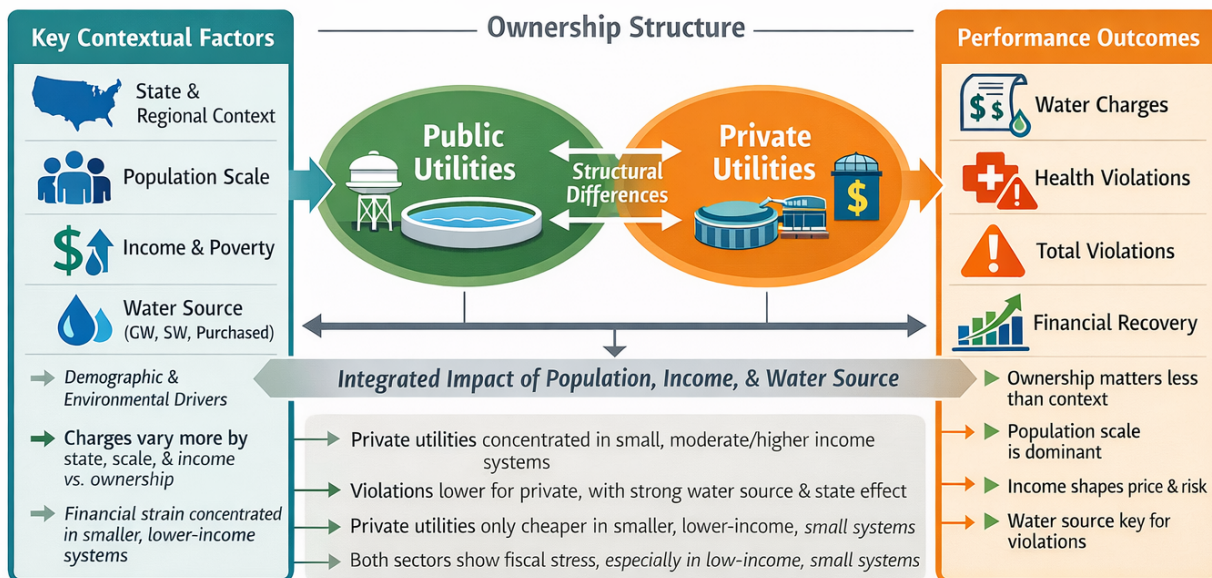
### 3.6 Random Forest Models

RF models were estimated for average water charges and 10-year health violations for each dataset (six RF models total). RF is a nonparametric ensemble method naturally capturing nonlinear relationships and multi-way interactions. All results are based on k-fold cross validation. Three interpretive tools are reported: (i) feature importance (MDI) ranks predictors by their contribution to reducing prediction error; (ii) SHAP values decompose each individual prediction into additive feature contributions, providing direction and magnitude; (iii) partial dependence plots (PDPs) reveal the functional form of each predictor's effect. Friedman H-statistics quantify interaction strength between predictor pairs. See Appendix C for method details.

**Note on RF violation models:** RF models are reported for 10-year health violations for all three datasets. Results for 5-year health violations, 5-year total violations, and 10-year total violations were also estimated and follow directionally consistent patterns. Those results are available from the authors on request.

## 4. Exploratory Analysis

Before formal modeling, we conducted an extensive exploratory analysis of all three datasets to understand distributional patterns, structural differences between public and private utility populations, and the likely nature of key relationships. This section summarizes the major findings and connects them to the formal modeling in Section 5. Figures not included in the main text are collected in the chapter appendix at the end of this section. The general structure of the exploratory analyses is illustrated below.



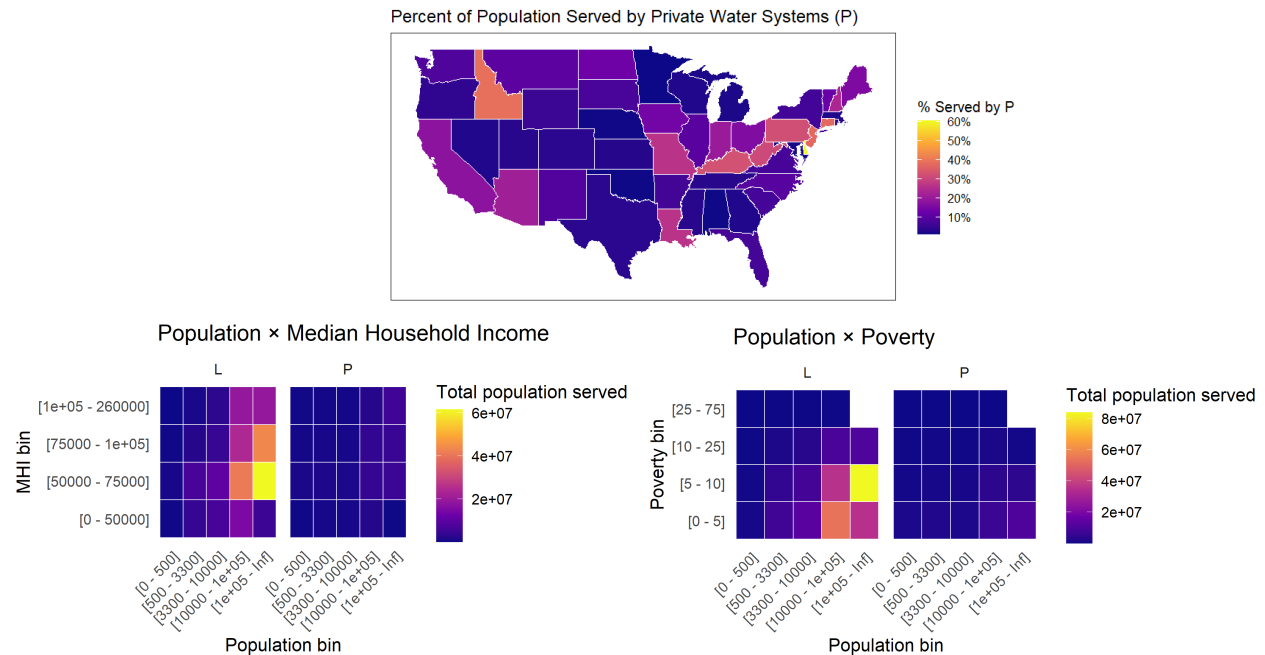
**Figure 1** Workflow for the Analysis. Data set creation and Processing to Modeling and Synthesis

Specifically, we explored how water charges, health violations, operating cost recovery, and related utility attributes vary with utility size, source water, and location. The intention was to get some understanding of the patterns in the data before formal regression and machine learning models were built, so that potential linear and nonlinear relationships could be anticipated and modeled.

The cross-cutting patterns: ownership, population scale, income, poverty, water source, and state context, align into a set of consistent takeaways about how public and private utilities differ, where they behave similarly, and which structural factors shape performance. In the sections that follow, we provide both the overall metrics (using Emma) and, where appropriate, the partitioning into the Public and Private data sets that are smaller than Emma due to the inclusion of the limited Financial data, so that the reader has a sense of the representative differences in the data sets used for analysis.

### 4.1 Who Serves Whom? The Structural Context of Ownership

Public and private utilities operate in systematically different demographic and geographic niches. Understanding this sorting is essential for interpreting differences in ownership and performance. We first look at the geographical distribution of utilities by ownership (L=Public, P=Private).



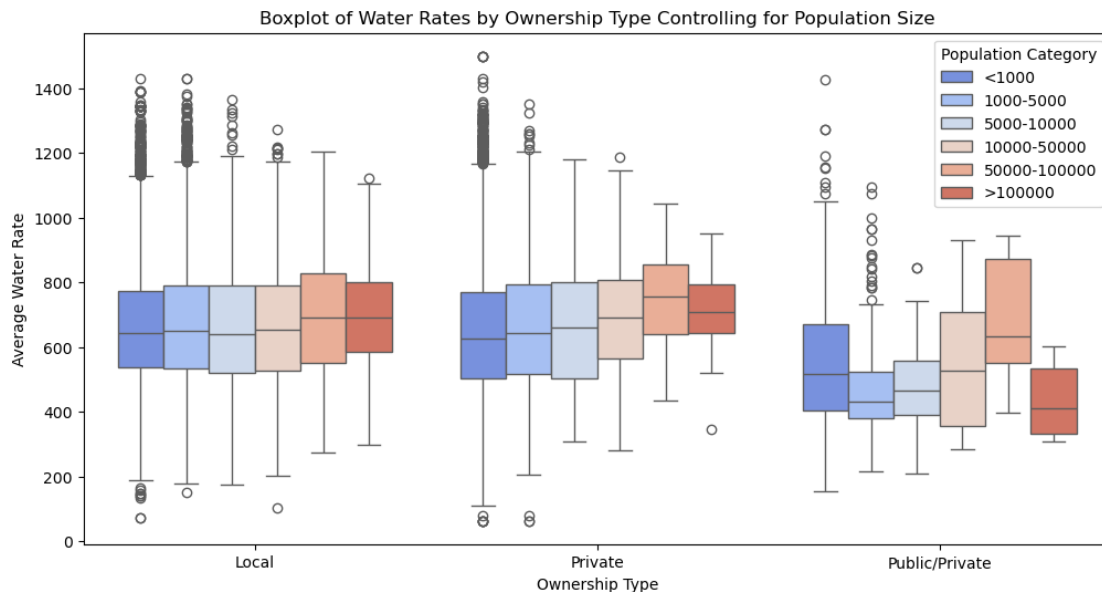
**Figure 2.** *Percent of Population served by Private Utilities by state, and total population served by community water system, population class, median household income, and percent poverty.* Public ownership is dominant. The largest total population served falls in the \$50k to \$75K annual MHI, with 5-10% poverty, and is served by public utilities. Private utilities typically serve much smaller population groups. The largest private utility counts (not shown) are in the \$50k to \$75k MHI range, in the 0-500 population bin across all income classes, and in the 0-5% poverty categories.

Figure 2 confirms the fundamental structural asymmetry: public utilities serve most of the population, particularly in the largest and most income-diverse communities. Private utilities are numerically concentrated in very small systems (population under 500) across all income groups, and many ultra-small private communities have the lowest poverty rates — consistent with investor-owned utilities targeting stable, lower-risk suburban and rural service areas where cost recovery is more predictable. This structural sorting has a major implication for interpreting raw ownership comparisons: privately owned utilities are often absent from the highest-risk, highest-cost demographic environments that public utilities are obligated to serve. The figure does not sufficiently highlight the “power law” structure for both private and public utilities, i.e., there is an exponential decrease in the number of utilities as the population they serve increases. The percentage of the population served by private utilities varies by state and, to a great extent, matches the states for which we were able to obtain financial data.

### Mississippi River Corridor — MRCTI Sub-Report

Mississippi River corridor states show a characteristic pattern: Louisiana, Mississippi, Arkansas, Tennessee, and Kentucky have low private market share and predominantly public utility service. Midwest corridor states also have limited private penetration, consistent with their Cluster 0 classification in the GLM state analysis.

## 4.2 Water Charges: Similar Medians, Strong Context Effects



**Figure 3.** Annual average water charges by ownership type (Local/Public, Private, Public-Private) and population category from the 2022 ACS. Median charges for public and private utilities are similar across population categories (~\$650–\$700/yr). Public-private utilities are systematically lower (\$450–\$600/yr).

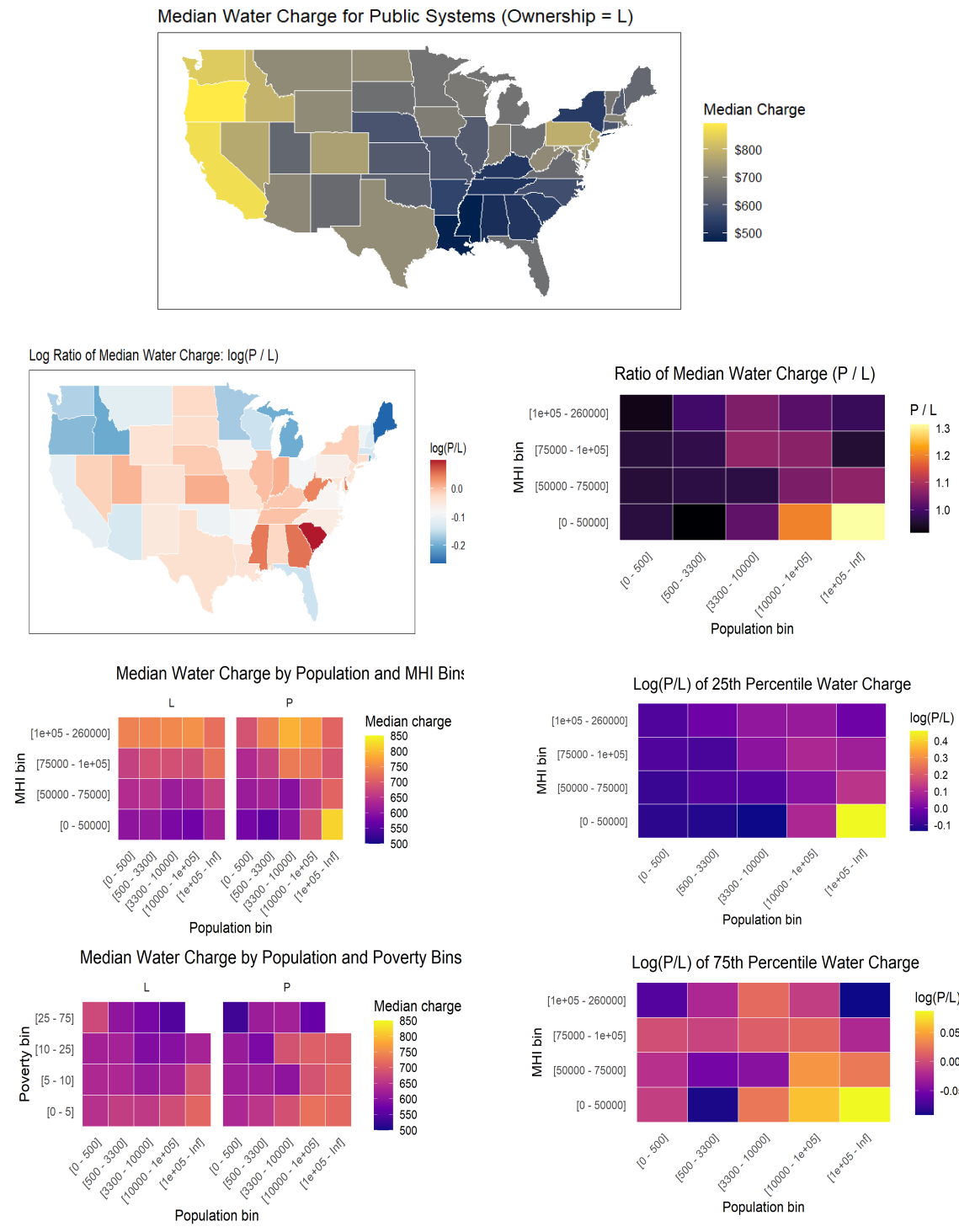
Figure 3 reveals the first key exploratory finding: **median annual water charges for public (L) and private (P) utilities do not vary substantially by population served — both center around \$650–\$700/year**. The overlapping distributions with wide boxes and many outliers in both categories indicate that ownership type alone is a poor predictor of what any individual customer pays. Public/private utilities typically have much lower water charges (\$450–\$600) than either of these two groups, but there are far fewer of them.

Figure 4 provides one of the most informative exploratory visualizations: the log ratio of private to public median charges in each income × population cell. **The pattern is strongly context-dependent. Private utilities charge less than public utilities in small systems (population quintiles Q1–Q2) across all income levels, but in larger systems serving low-income communities (Q4–Q5 population, Q1–Q2 income), private utilities charge more.** This regime-switching is cheaper on a small scale but more expensive in large low-income systems, which directly motivated the use of Random Forest models in Section 5. The private utility-to-public utility charge ratio is highest in the lowest-income and highest-population-served bins, and lowest in the smallest-population bins at any income level. The lowest 25<sup>th</sup> percentile of water charges shows considerably more variation between public and private utilities than the upper 75<sup>th</sup> percentile, but the pattern is generally similar. The *private-public charge differential narrows substantially at the 75<sup>th</sup> percentile of charges*. The ownership gap is primarily in the middle of the charge distribution, not at the extremes. **High-bill customers face similar charges regardless of ownership type — it is at moderate usage levels and in specific demographic contexts where the ownership difference is most consequential.**

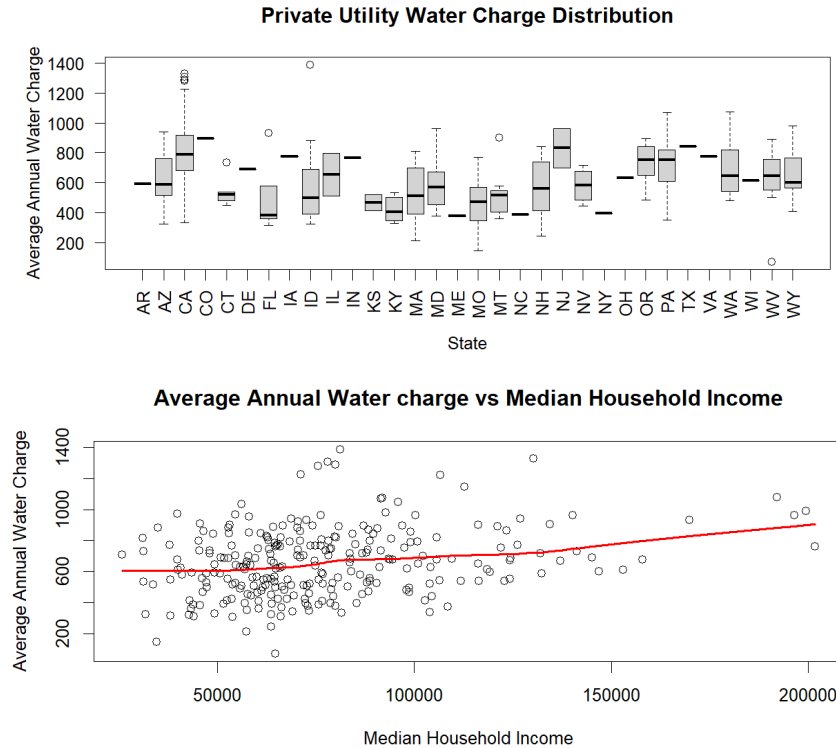
Figure 5 shows considerable variation in water charges within and across states for private utilities. We also see that charges generally increase with median household income. However, the highest charges are observed among lower-middle- and middle-income households. The patterns for public utilities (not shown) are similar.

Figure 6 shows that the geographic coherence of states by Pricing, Income, and Ownership is notable — the exploratory k-means clustering recovers similar regional patterns that emerge from the formal GLM state fixed effects in Section 5.

The *standard deviation of charges within a utility* (charge variability) is a novel outcome variable that has not been analyzed in the literature. We noted that charge variability increases substantially with both income and population served. This is consistent with two mechanisms: larger utilities with diverse customer classes generate more within-utility charge spread; income-diverse service areas are more likely to have tiered or income-graduated rate structures. High charge variability does not necessarily indicate unfair pricing — it may reflect progressive rate design. The formal models in Section 5 confirm that charge variability is tightly coupled to average charge levels (partial  $R^2$  of charge variability predicting average charge  $\approx 43\%$  in the public dataset).

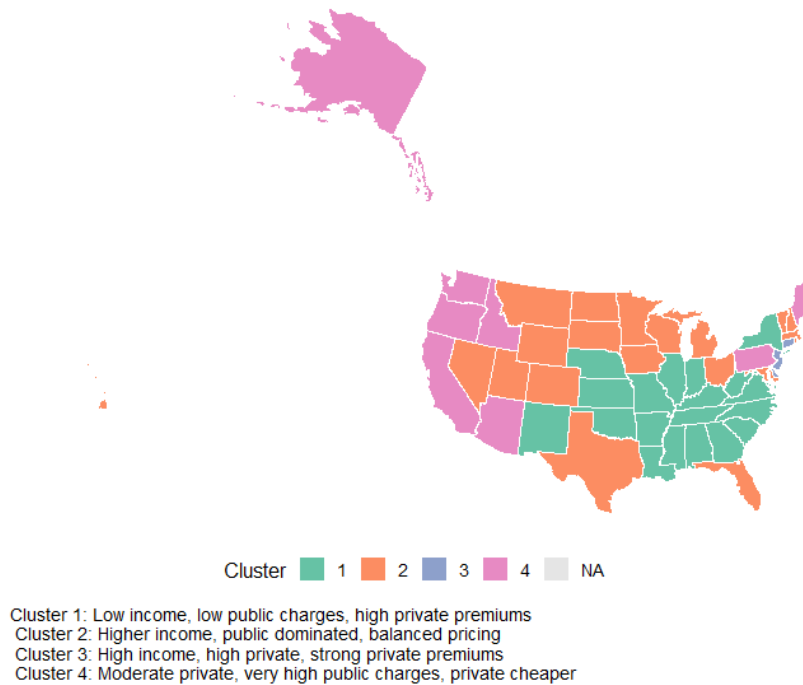


**Figure 4. Statewide variation in Median Water Charge for Public Utilities, The ratio of Privat to Public Utility charges and demographics of water charges by population served, MHI and Poverty %.** Red = private more expensive than public; blue = private cheaper. The pattern is strongly context-dependent: private utilities are cheaper in small systems but more expensive in large, low-income systems. At the 75th percentile (high-bill customers), differences between public and private utilities largely disappear.



**Figure 5. Distribution of Average Annual Water Charge by State and by Median Household Income for Private Utilities.** The variation by state is substantially higher than by population class. A similar pattern emerges for public utilities. Note also the substantial variation within a state. Note the trend for increasing charges as the median household income increases.

### State Clusters Based on Ownership, Pricing, and Income

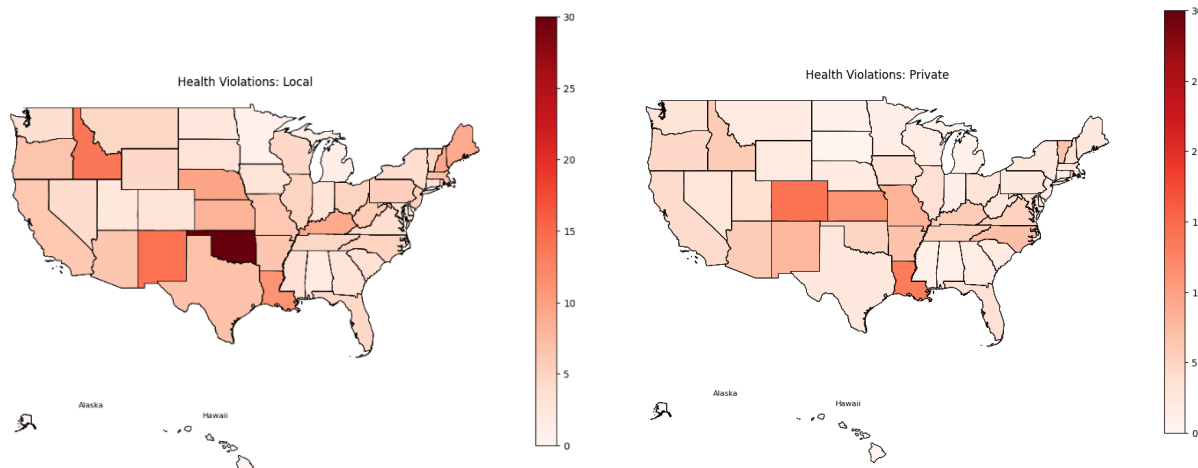


**Figure 6. State clusters based on ownership share, median household income, and median water charge (k-means, 4 clusters).** Cluster 1 (teal): Low income, low public charges, high private premiums — Deep South pattern. Cluster 2 (orange): Higher income, public-dominated, balanced pricing. Cluster 3 (purple): High income, high private share, strong premiums. Cluster 4 (pink): Moderate private, very high public charges, West Coast/Alaska.

### Mississippi River Corridor — MRCTI Sub-Report

Mississippi River corridor states cluster primarily in Cluster 2 (WI, IA, IL, MN, MO) and Cluster 1 (LA, MS, TN, KY). Arkansas is a notable exception, clustering with higher-price states despite its Deep South demographics — an anomaly explored further in the companion sub-report.

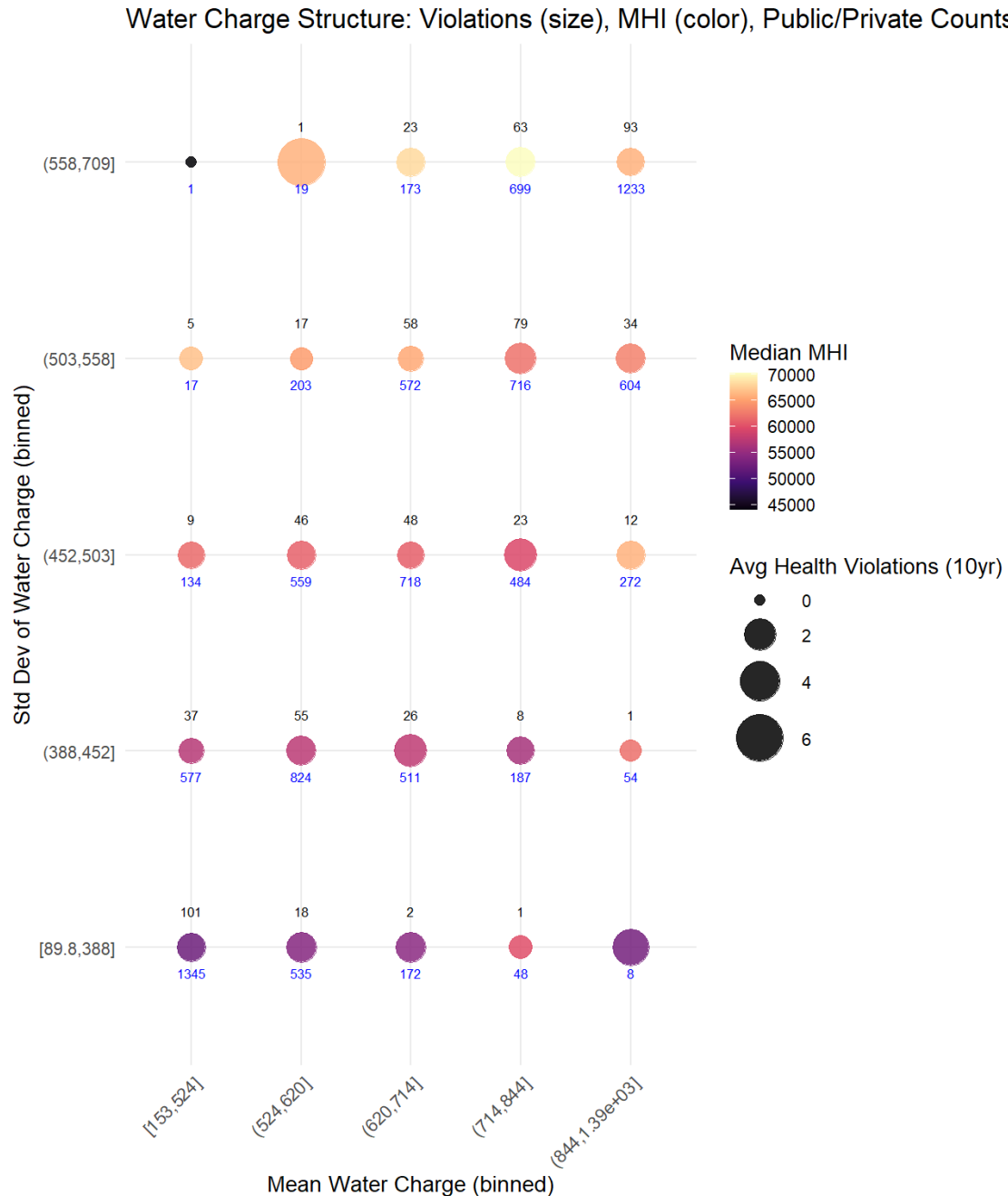
## 4.3 Health Violations: Ownership, Source, and Geography



**Figure 7. 10-year Health Violations by ownership.** Note that a few states dominate the patterns and that some of these states are common across ownership categories.

The health violations for 2015-2024 by ownership are shown in Figure 7: the violation burden is concentrated in a small number of states (Pennsylvania, New Jersey, California, Arizona). This geographic concentration is an important context for the RF finding in Section 5 that private utility violations are primarily explained by financial structure (PCF2 and PCF1) rather than state identity — the state pattern in raw data is largely explained by the financial characteristics of utilities in those states.

A multi-variable visualization of 10-year health violations relative to mean water charges, the standard deviation of water charges within a utility, and the median household income is presented in Figure 8. High violations with a low mean and high standard deviation of charge correspond to a category with 1 private and 19 public utilities, accounting for 0.1% and 0.2% of the utilities in the sample (for which financial data is available). These also correspond to higher-median-household-income blocks. **The lowest standard deviation of water charges corresponds to the lowest median household income group, which typically has lower mean water charges and higher violations.** The increase in the standard deviation of water charges with income, could be due to differences in water use or to subsidies. **Violations increase with middle categories of mean water charge and standard deviation of water charge.** Median household income does not seem to correspond to either water charges or violations.



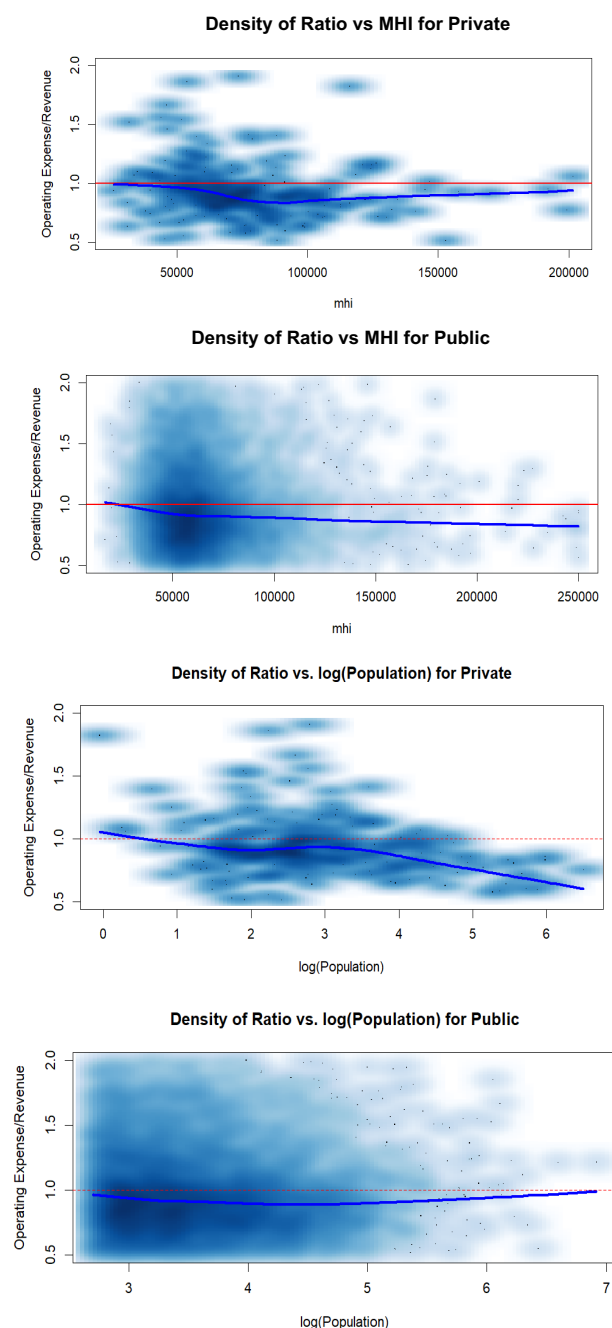
**Figure 8.** 10-year Health Violations vs Average and Standard Deviation of Water Charge parametric to median MHI. The number of private (black counts) and public (blue) utilities in each bin is listed above each bubble to illustrate the sample sizes that indicate the joint variation in the charge structure, mhi and health violation outcomes.

#### 4.4 Financial Performance: Cost Recovery and the Role of Taxes

Cost recovery (Operating Expense-to-Revenue Ratio) has been a recurring concern for water utilities. We explore this statistic in Figures 9 and 10.

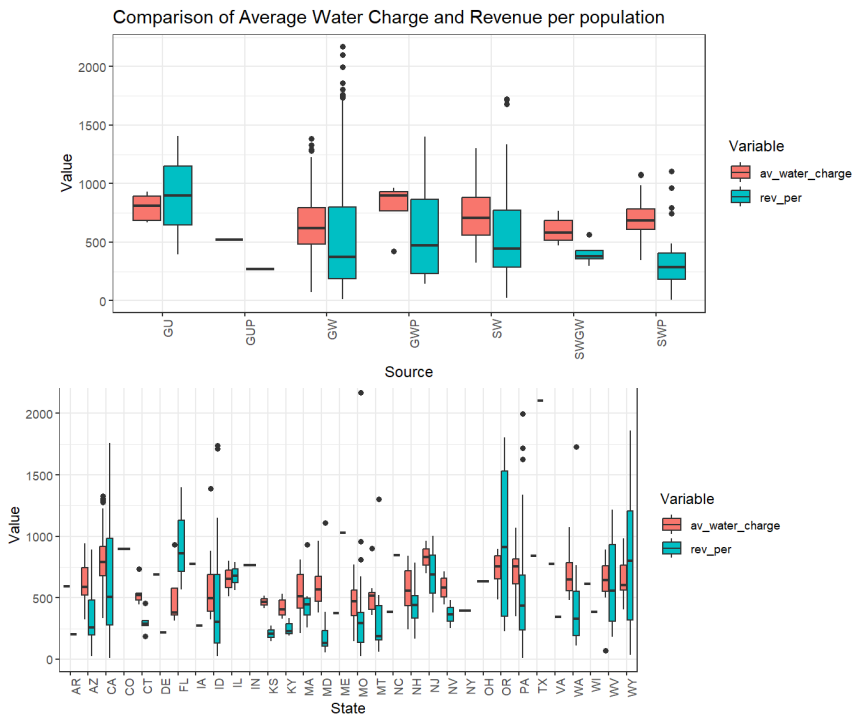
A consistent trend for both ownership types is that cost recovery improves with larger population size and higher MHI. The population effect is much greater for private utilities than for public ownership. These results are consistent with expectations for the most part.

**Figure 9.** Expense Revenue ratio for Public and Private utilities for the 2022 Financial Data. The lower panels show how Expense revenue ratio changes as a function of median household income and population for private and public ownership. The private data is from annual financial reports, while the public data is from the Municipal Financial Survey and uses water utility financial reporting. A ratio > 1 means total operating expenses exceeded reported annual revenue in 2022.

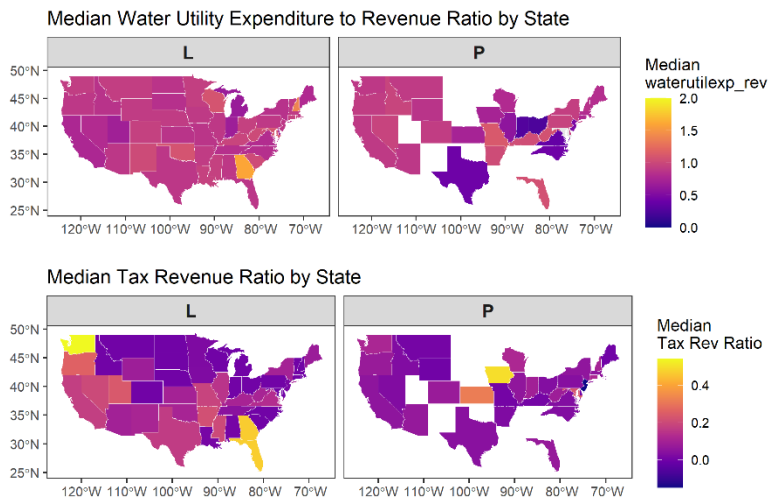


An interesting observation is that the MHI effect is much smaller than the population effect for private utilities, whereas it is somewhat higher for public utilities. Spot checks of large public utilities with operating expense-to-revenue ratios greater than 1 indicate that they typically do not correspond to the highest MHI categories. Similarly, the private utilities with the lowest ratios and higher populations belong to middle-income communities. The highest cost recovery (25<sup>th</sup> percentile and low ratios) corresponds to the largest private operations with middle-to-high MHI. The worst (75<sup>th</sup> percentile and high ratios) are for the private, small, and richest group, which is curious. Private has a higher cost recovery in all MHI and population categories.

**Figure 10. A Dive into Private Utility Cost Recovery Drivers – Comparing Per Capita Revenue and Average Water Charge by Water Source, by State.** Note the wide variation in water charges and revenue per capita, with no consistent pattern by source or state but we see some anomalies with relatively high taxes in KS.



**Figure 11. Comparing Median Expense revenue ratio and Tax to Revenue Ratio by State for Public (L) and Private (P) utilities.** Significant contrasts in public vs private cost recovery are notable at the state level. Public utilities typically do not have taxes but they seem to report quite high taxes in some states, e.g., WA and FL



## 5. Statistical Inference: GLM and Random Forest Results

This section presents integrated results from GLMs and RF models across all three datasets and both primary response variables. The analysis is organized by dataset tier. For each dataset, GLM findings establish statistically confirmed linear associations; RF findings reveal functional form, interactions, and nonlinear relationships that the GLM cannot detect. A cross-method comparison closes each sub-section. We refer the reader to Appendix C for the methodological details of the GLM and RF models used here, including caveats and interpretations. Here we will present the results from each method, organized by data set. The GLM results are presented first, followed by the RF results. A synthesis of the findings for each section is presented at the end of the section, with a final synthesis of the conclusions from both sets of Analyses. Please refer to Appendix D and Appendix E for the detailed results

### 5.1 EMMA Dataset: Full Utility Population without Financial Data

We focus on the relationship between ownership structure and two central performance dimensions of water utilities: average annual water charges and 10-year health violations. Using Generalized Linear Models with Lasso regularization with parameters selected by cross-validation, we estimated conditional associations while accounting for water source, system scale, state effects, and community socioeconomic structure. Of these, variations across states account for the dominant variability in the covariates. The results reveal a clear asymmetry between financial and health outcomes.

- **Ownership structure is strongly and consistently associated with average water charges.** After controlling for demographic, operational, and structural characteristics, multiple ownership categories exhibit large and statistically significant positive effects on pricing. *The average water charge for publicly owned utilities is \$41 per year higher than for privately owned utilities.* These differences are economically meaningful and robust to high-dimensional adjustment. This indicates that governance structure is materially associated with rate levels, even when accounting for underlying contextual factors.
- In contrast, **ownership is not a strong or systematic predictor of long-run health violations once water source, scale, and socioeconomic characteristics are provided.** Violation frequency appears to be more strongly driven by environmental and operational complexity (particularly surface-water dependence) and by community structural conditions. Only limited ownership effects remain statistically significant in the violations model.
- Across both regressions, **socioeconomic structure and system scale emerge as consistent and important predictors.** These results underscore the importance of community context and infrastructure conditions in shaping both pricing and compliance outcomes. Larger systems have lower violations and average annual water costs. *Systems with a higher population relative to the number of service connections have higher water charges and more violations.*
- Additionally, the *positive association between violation burden and pricing suggests that regulatory stress may translate into higher consumer costs*, reflecting remediation, capital investment, or compliance expenditures.

Overall, the findings support *four central conclusions*:

1. *Differences due to state are larger than differences due to other factors.*
2. *Ownership differences are more pronounced in financial outcomes than in compliance outcomes.*
3. *Water source and community socioeconomic structure are fundamental determinants of both pricing and health performance.*
4. *Compliance burden and pricing are linked, suggesting financial feedback from regulatory conditions.*

These results advocate for a more nuanced view of ownership effects—one that distinguishes governance-related pricing differences from structurally driven compliance outcomes.

The regression coefficients for the variables retained by Lasso cross-validation and the statistical significance (p-values<0.05) of the hypothesis test for no effect are presented in the table below. The categorical variables (source, state, ownership) are scored relative to a “reference” category, which for all these results is Alaska for State, Groundwater under the influence of surface water for source, and federal for ownership.

### 5.1.1 Average Water Charges & 10 year Health Violations— EMMA

The EMMA charges GLM (N=43,532, Gaussian link,  $R^2=0.23$ ) identifies state effects as the dominant source of explained variance (partial  $R^2=19.9\%$ ), followed by demographics and ownership. Table 3 summarizes key GLM coefficients and partial  $R^2$  values. *Note: all coefficients are for predictors significant at the 5% level unless otherwise stated; state coefficients are expressed relative to Alaska as the reference state.*

**Table 3.** GLM results for EMMA average water charges and 10-year health violations. Coefficients statistically significant at 5% level. State effects relative to Alaska (reference).

| Predictor                              | Charge (\$/yr) | 10-yr health viol. | Interpretation                                                |
|----------------------------------------|----------------|--------------------|---------------------------------------------------------------|
| Source: Groundwater (GW)               | -35.65         | -0.38              | GW is cheaper and safer than the GW-under-influence reference |
| Source: Surface Water (SW)             | +5.58          | +2.00              | SW is more expensive and has substantially more violations    |
| Source: GW Purchased (GWP)             | -27.07         | -0.28              | Purchased GW is cheaper and safer                             |
| Ownership: Public/Local (L)            | +103.85        | -0.07              | Public utilities charge more; marginal violation reduction    |
| Ownership: Private (P)                 | +62.98         | -0.13              | Private charges less than public (~\$41/yr net diff.)         |
| Scale PCS1 (system size)               | -2.26          | -0.10              | Larger systems: lower charges and fewer violations            |
| PCD1 Economic Affluence                | +5.36          | -0.10              | Affluent areas: higher charges, slightly fewer violations     |
| PCD2 Wealthier, Urban Density & Rental | +19.34         | -0.15              | Urban/diverse areas: higher charges, fewer violations         |
| PCD4 Deep Poverty Gradient             | -4.07          | +0.08              | Deep poverty: lower charges, more violations                  |
| 10-yr health violations                | +0.85          | —                  | Each additional violation → \$0.85/yr higher charge           |
| Partial $R^2$ : State                  | 19.9%          | 7.3%               | Dominant factor in both models                                |
| Partial $R^2$ : Source                 | 0.4%           | 2.7%               |                                                               |
| Partial $R^2$ : Ownership              | 1.3%           | 0%                 |                                                               |
| Partial $R^2$ : Demographics           | 1.5%           | 0.4%               |                                                               |
| Partial $R^2$ : Total                  | 23.1%          | 10.5%              |                                                               |

The EMMA GLM confirms six primary findings:

*First*, state effects dominate: state partial  $R^2$  (19.9% for charges, 7.3% for violations) exceeds every other predictor group. California utilities charge the most; Mississippi utilities charge the least — and these state effects persist after controlling for ownership, source, demographics, and scale.

*Second*, ownership matters for charges but not violations: public utilities charge approximately \$41/yr more than private (103.85 – 62.98, both relative to federal reference). Ownership has minimal independent effect on violations once other factors are controlled.

*Third*, groundwater sourcing is consistently protective — \$35.65 cheaper and 0.38 fewer violations per decade relative to the groundwater-under-surface-water-influence reference.

*Fourth*, Economic Affluence and urban diversity both predict higher charges and fewer violations.

*Fifth*, system scale is protective: larger systems have lower charges and fewer violations.

*Sixth*, violations and charges are positively associated: each additional 10-year health violation corresponds to \$0.85/yr in higher charges, consistent with compliance costs feeding through to ratepayers.

### RF findings: EMMA charges

The RF charges model (CV  $R^2=0.40$ ) substantially outperforms the GLM ( $R^2=0.23$ ) and adds three critical nonlinear findings: (1) *PCD2 (Wealthier, Urban Density & Rental ) exhibits a saturating sigmoid — the marginal charge premium from additional urban wealth diminishes above a threshold, so the GLM's linear coefficient overestimates the effect at extreme urban values;* (2) *the PCD1 × PCD2 interaction is superadditive — communities combining Economic Affluence with urban diversity face charges higher than the sum of individual effects (+25 units above additive prediction at the joint extreme);* (3) *California's mean SHAP reaches +60, Mississippi's –60 — confirming and quantifying the GLM's state dominance finding. The RF reveals that most of the top 12 predictors (PCD1 through PCD8, state indicators, and scale PCs) exhibit nonlinear PDPs including threshold-then-ramp (PCD1), saturating sigmoid (PCD2), and inverted-U (PCD3) shapes.*

### 5.1.2 10-Year Health Violations — EMMA

The EMMA violations GLM (Poisson link, pseudo- $R^2=0.105$ ) is structurally different from the charges model. State effects remain dominant (7.3% partial  $R^2$ ), but source type takes second place (2.7%). Oklahoma and Louisiana have the largest positive state coefficients for violations. Groundwater source reduces violations (–0.38); surface water increases them (+2.00). System scale (PCS1) is protective (–0.10). Economic Affluence (PCD1, –0.10) and urban diversity (PCD2, –0.15) both reduce violations.

The RF violations model (CV  $R^2=0.20$ ) confirms Oklahoma's dominance — state\_OK ranks first in feature importance with MDI approximately 4× larger than any other feature. *The RF adds a critical interaction finding: primary source code is the dominant interaction variable, with a Friedman H-statistic of ~0.200 — the largest interaction measure across all eight models in this series, approximately 8× larger than the next feature.*

**Surface water × poverty is multiplicative, not additive: surface water utilities in high-poverty communities average up to 7× more violations than groundwater utilities in affluent communities.** The GLM's separate linear coefficients (+2.0 for surface water, negative for poverty) describe the average of a 1× to 7× range.

#### Mississippi River Corridor — MRCTI Sub-Report

In the EMMA model, Louisiana and Oklahoma are the two largest positive violation states after controlling for all covariates. Mississippi has the largest negative charge SHAP — the cheapest state nationally after controlling for factors. Missouri shows mild positive violation effects, driven by surface-water river-basin exposure.

## 5.2 Public Utilities

### 5.2.1 Average Water Charges — Public

The public utility charges GLM (N=8,007,  $R^2 \approx 0.62$ ) achieves substantially higher explanatory power than the EMMA model, reflecting the addition of municipal financial variables. The *standard deviation of charges* (charge variability) is included as a predictor and has the largest partial  $R^2$  of any individual predictor (~43%), consistent with a strong mutual relationship between rate level and rate heterogeneity. State effects remain at 19.6% partial  $R^2$ .

The GLM identifies four state clusters based on charge deviations and violation performance:

- **Cluster 0 — Low-Price, Structurally Neutral (Baseline Majority):** Most Midwest and Southeast states. Strongly negative average charge deviations, moderate compliance, no extreme fiscal stress.
- **Cluster 1 — Persistent High Health Violations:** Oklahoma, Louisiana, New Mexico. Very high 5- and 10-year health violations; pricing not proportionally higher.
- **Cluster 2 — High Total Violation Burden:** Arizona, West Virginia, Pennsylvania. Elevated long-run total violations, moderate-to-high health violations.
- **Cluster 3 — Higher-Price / Financially Stable:** California, Washington, New Jersey, New York, Michigan, Vermont, Oregon, Maryland, and notably Arkansas. Less-negative or positive charge deviations, lower long-run violations, stronger financial stability.

**Table 4.** Selected GLM coefficient patterns for public utilities. All coefficients statistically significant at 5% level.

| Predictor                            | Charge effect  | Violation effect | Interpretation                                                                                  |
|--------------------------------------|----------------|------------------|-------------------------------------------------------------------------------------------------|
| <b>State: Washington, CA, OR</b>     | +20 to +45     | Varies           | West Coast states charge substantially more; largest positive state premiums nationally         |
| <b>State: Mississippi, Louisiana</b> | -15 to -30     | LA positive      | Deep South discount; LA simultaneously below-market charges and above-average violations        |
| <b>State: Oklahoma (violations)</b>  | Moderate -     | Largest +        | Oklahoma public utilities have dramatically more violations than any other state after controls |
| <b>PCD1 Economic Affluence</b>       | + (continuous) | — (negative)     | Higher poverty → higher charges AND more violations: compound disadvantage                      |
| <b>PCF7 Debt Leverage</b>            | + step at 0    | + moderate       | High debt relative to revenue: +\$46 charge premium (RF step-function) and more violations      |
| <b>Primary source SW</b>             | +              | ++ strong        | Surface water: higher charges, dramatically more violations                                     |
| <b>Opex &gt; Revenue (violation)</b> | N/A            | + positive       | Financial stress → more violations; compliance-cost feedback confirmed                          |

The RF public charges model produces several findings with no GLM equivalent:

- **The poverty-charge paradox is confirmed and quantified.** PCD1 (Economic Affluence) is the fourth-ranked RF predictor and exhibits a continuously increasing, monotone relationship: public utilities serving the poorest communities charge more, not less. The range spans approximately \$60 across the PCD1 support — the largest marginal range of any continuous demographic feature in the public charges model. The relationship accelerates above the mean poverty level.
- **Debt leverage (PCF7) is a binary threshold effect.** The RF identifies a step-function: utilities above the debt-to-revenue threshold charge approximately \$46/yr more — the largest single financial effect

across all eight models. The GLM captures this as a positive linear coefficient; the RF reveals the entire effect is concentrated at the threshold.

- **PCD1 × PCD3 superadditive interaction.** *Communities simultaneously poverty-concentrated and health-insurance-deficient face predicted charges approximately \$25 above what additive effects would suggest* — the most important compound-disadvantage finding in the public utility analysis.

## 5.2.2 10-Year Health Violations — Public

The public violations GLM (Poisson, pseudo- $R^2 \approx 0.17$ ) is dominated by source type and state, with Oklahoma, Louisiana, and Arizona showing the largest positive state deviations. Surface water predicts more violations. The GLM finds a positive association between operating expense pressure and violations — the compliance-cost linkage.

The RF public violations model (CV  $R^2 \approx 0.25$ ) produces its most important finding in the source × poverty interaction. The Friedman H-statistic for primary source code ( $\sim 0.200$ ) is the single largest interaction measure observed across all eight models — approximately 8× larger than the next feature (PCD1 at  $\sim 0.025$ ). The 2D partial dependence plot for source × PCD1 shows predicted violations ranging from approximately 1.0 (groundwater + affluent area) to 7.0 (surface water + extreme poverty) — a 7-fold range that the GLM's separate linear coefficients dramatically understates.

A second major RF finding concerns financial balance: the PCFS1 × PCFS8 interaction shows a non-monotone U-shaped compliance response around the revenue-efficiency breakeven. Violations are highest at both revenue deficit extremes AND revenue surplus extremes, with minimum violations at the breakeven point. The GLM captures only the deficit side (positive association between expense pressure and violations), missing the surplus-side compliance risk entirely.

### Mississippi River Corridor — MRCTI Sub-Report

Oklahoma's public violation state effect (MDI  $\approx 0.29$  in the RF, SHAP reaching +2.5) is qualitatively different from any other state — a structural, geological, and regulatory phenomenon specific to its public sector. Louisiana's moderate positive violation SHAP (+0.3–0.5) combined with its large negative charge SHAP (–15 to –20) makes it the national compound-disadvantage case study: below-market charges coinciding with above-market violations in public utilities. Private utilities in Louisiana show no elevated violation state effect.

## 5.3 Private Utilities

### 5.3.1 Average Water Charges — Private

The private utility charges GLM ( $N=277$ ,  $R^2 \approx 0.63$ ) achieves high explanatory power despite the small sample, with demographics explaining the largest partial  $R^2$  (31%). State effects are mostly non-significant; only a handful of states (KS, KY, NC, NY) show significantly lower charges, and Kentucky's  $\sim \$400$  private charge discount is one of only four significant state charge effects nationally.

The RF private charges model (CV  $R^2 = 0.24$ ) reveals a fundamentally different pricing architecture from both the EMMA and public models: virtually every predictor exhibits a step-function (binary regime) partial dependence plot. Key findings:

- **California and Pennsylvania create high state premiums.** California and Pennsylvania dominate the SHAP magnitude map for private utility charges. These state effects likely encode regulatory rate-of-return frameworks and market structures specific to investor-owned utility sectors in these states.

- **Knowledge economy concentration activates a discrete premium tier.** PCD3 (Knowledge Economy) shows the largest marginal range (+36 units) and activates sharply at the extreme upper tail — likely Silicon Valley and Research Triangle-type markets. Below that threshold, the knowledge economy gradient provides minimal pricing differentiation.
- **Tax-dominant financial strategy elevates charges (PCF2).** PCF2 (Tax Impact) shows a pure binary step function: utilities where tax burden dominates the financial mix (PCF2 > 0) charge approximately \$20.50 more per year than utilities with below-average tax burden.

**Table 5.** Feature importance ranking: RF private utility average water charges. Partial dependence shape describes the functional relationship between the feature and predicted charges.

| Rank | Feature                                | Category              | PDP shape                | Interpretation                                                         |
|------|----------------------------------------|-----------------------|--------------------------|------------------------------------------------------------------------|
| 1    | state_CA                               | Geographic            | Consistent high +SHAP    | California creates largest irreducible charge premium                  |
| 2    | PFS1 Per-Capita Affluence              | Financial (scaled)    | Step at minimum          | Binary: financially active utilities charge more                       |
| 3    | PCD3 Knowledge Economy                 | Demographic           | Step at +1.8 (+36 units) | Largest demographic range; premium only at extreme tech-cluster values |
| 4    | PCD1 Economic Affluence                | Demographic           | Step at 0 (+23 units)    | Binary affluence discriminator: above/below mean                       |
| 5    | PCF2 Tax Impact                        | Financial (aggregate) | Pure step at 0 (+20.5)   | Tax-dominant utilities charge \$20.50 more — binary discriminator      |
| 6    | PCF1 Financial Scale                   | Financial (aggregate) | Step at minimum (+14)    | Financial viability threshold; above it, scale adds nothing            |
| 7    | PCD2 Wealthier, Urban Density & Rental | Demographic           | Step at -1.8 (+27 units) | Two-regime pricing: urban-diverse vs. non-urban                        |

### 5.3.2 10-Year Health Violations — Private

The private violations RF model (train  $R^2=0.19$ , CV  $R^2=0.06$ ) has substantially lower explanatory power than any other model in the series, reflecting the inherently difficult-to-predict violation behavior of the small private utility population (N=242 companies). Despite low overall  $R^2$ , the model identifies a clear structural signal:

- **Financial structure dominates; state identity does not.** PCF2 (Tax Impact) and PCF1 (Financial Scale) are the top two predictors by a wide margin. State indicators are essentially absent from the top 20 — only Pennsylvania shows a detectable state SHAP effect. This contrasts sharply with public utility violations where state\_OK dominates at MDI $\approx$ 0.29.
- **PCF2's binary step is the largest single violation effect across all eight models.** Tax-dominant private utilities (PCF2 > 0) average 2.34 more violations per decade than opex-dominant utilities — approximately doubling the baseline violation count. The same PCF2 sign that predicts higher charges in the charges model also predicts more violations here. Tax-dominant investor-owned utilities therefore impose a compound burden: above-market charges AND worse compliance.
- **PCF1's indication was discounted after checking the raw data. A few cases of acquisitions of public utilities (that were small and in poorer areas) that had high prior violations by larger private utilities biased the results vs financial scale. Models that did not include these cases did not produce the same results. Consequently, we dropped the conclusions associated with PCF1 from consideration.** This is less of a problem with the EMMA data set where the analysis is at the PWSID level and hence the small, poor attributes at the PWSID level accounted for the correct

relationships with scale. Sensitivity analyses with respect to other predictors did not lead to a similar problem.

The GLM private violations model identifies operating margin (PCF3) as the dominant predictor with large positive coefficients for both 5-year and 10-year health violations, consistent with the RF finding that financial structure dominates private utility violations.

## 5.4 Operating Deficit (Financial Performance) Models

The GLM logistic models for the operating deficit indicator (operating expenses exceeding revenue) provide additional insight on financial resilience:

- **Georgia stands out for public utilities.** Georgia has the largest positive state coefficient for operating deficit odds in the public model, suggesting structural municipal fiscal stress beyond what financial PCs can explain. Midwest and Mountain West states show generally lower deficit odds.
- **Financial structure PCs drive private sector deficit risk.** The private deficit model is concentrated in financial structure PCs contrasting operating expenses with taxes and revenues.
- **Deficit risk is associated with more violations.** The positive association between operating expense pressure and violations in both EMMA and public GLMs confirms a compliance-cost feedback: financially stressed utilities compromise maintenance and monitoring, leading to more violations. This reinforces the policy priority of preventing financial deterioration before it cascades into compliance failure.

## 5.5 Public vs Private Assessment across response variables for significant covariates

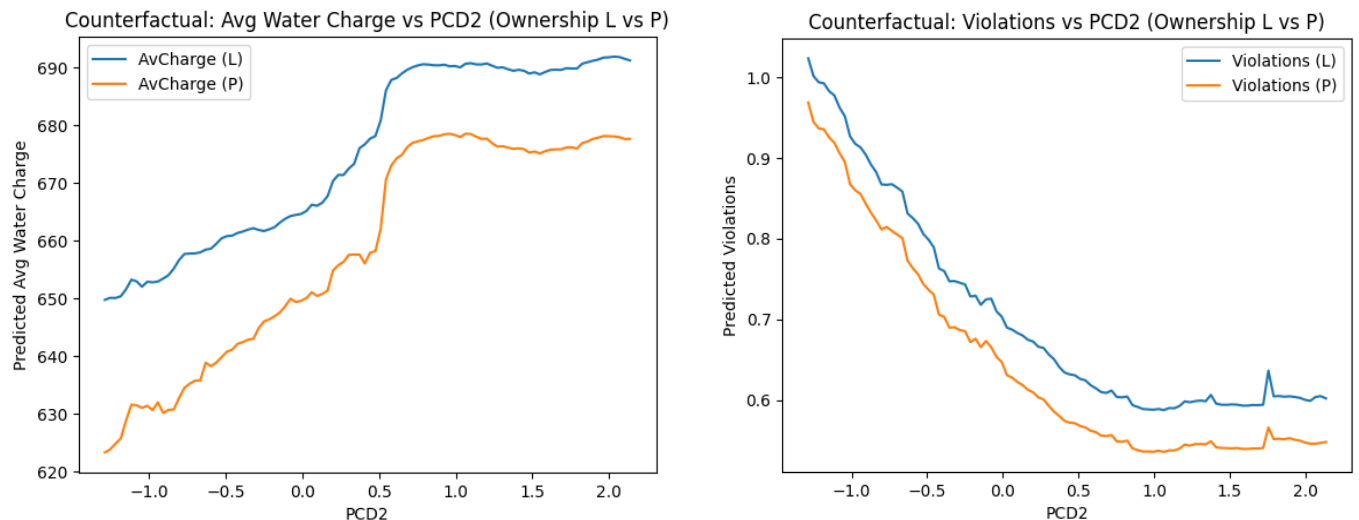
An aggregate summary of the importance of different predictors across the water charge and violations variables is presented in Table 6. It is important to note that we used PCA to decorrelate the covariates and then included all covariates in the model, i.e. their significance accounts for the effect of all other variables simultaneously modeled. We note some striking differences:

1. State is generally important in all cases except for health and total violations for Privates. This suggests that while state regulatory efforts affect pricing, and effectiveness of regulation for public utilities, the private utilities as a group have compliance behavior that is not state dependent.
2. Water Source varies in importance for charges for both public and private, and for violations for public utilities, but does not emerge as a determinant of violations for the private utilities, further adding to the argument that private utilities compliance efforts seem to be more robust to geography and source.
3. Interestingly, population seems to be related to violations but not to charges across both groups. Lower populations tend to have higher violations.
4. The demographic variables tend to have the highest effect on water charges and their heterogeneity and on violations for the public utilities.
5. The utility financial variables typically affect all response variables. PCF4 which measures the expense gap between operating and total expenses seems to be a strong determinant of private sector average charge and also its heterogeneity. It does not influence violations.

***Figure 1 presents an interesting observation that conditional on all other variables being used, the marginal dependence of average water charges and health violations on PCD2 (Wealthier, Urban Density & Rental ) shows a distinct performance advantage for private utilities over public utilities in that they appear to price lower and have lower violations across all demographic groups. Charges increase and violations decrease as urban wealth increases up to a point, and are constant after that point.***

**Table 6** *Qualitative Assessment across response variables and models.* The scores range from 0 to 3, where 0 means that the covariate is not significant and 3 means that it is highly significant and explains more variance.

| Covariate                        | Public Av | Private Av | Public Sd | Private Sd | Public 5y H | Private 5y H | Public 10y H | Private 10y H | Public 5y T | Private 5y T | Public 10y T | Private 10y T |
|----------------------------------|-----------|------------|-----------|------------|-------------|--------------|--------------|---------------|-------------|--------------|--------------|---------------|
| state                            | 3         | 3          | 3         | 3          | 3           | 0            | 3            | 0             | 2           | 0            | 3            | 0             |
| primary_source_code              | 1         | 3          | 1         | 2          | 2           | 0            | 2            | 0             | 1           | 0            | 1            | 0             |
| population                       | 0         | 0          | 0         | 0          | 1           | 3            | 1            | 3             | 1           | 0            | 1            | 2             |
| PCD1                             | 2         | 3          | 2         | 3          | 1           | 0            | 1            | 0             | 1           | 0            | 1            | 0             |
| PCD2                             | 1         | 2          | 2         | 3          | 1           | 0            | 1            | 0             | 1           | 0            | 1            | 0             |
| PCD3/4/6/7<br>(labor/poverty)    | 1         | 2          | 2         | 2          | 1           | 0            | 1            | 0             | 1           | 0            | 1            | 1             |
| PCD10 (high middle               | 0         | 1          | 0         | 2          | 0           | 2            | 0            | 0             | 0           | 0            | 0            | 0             |
| PCD19 (wealthy,<br>private only) | 0         | 0          | 0         | 1          | 0           | 0            | 0            | 0             | 0           | 0            | 0            | 0             |
| PCF1                             | 1         | 1          | 1         | 0          | 1           | 3            | 1            | 3             | 1           | 1            | 1            | 2             |
| PCF2                             | 0         | 1          | 1         | 0          | 1           | 2            | 1            | 2             | 0           | 2            | 0            | 1             |
| PCF3                             | 1         | 1          | 1         | 1          | 1           | 3            | 1            | 3             | 1           | 1            | 1            | 1             |
| PCF4                             | 0         | 3          | 1         | 3          | 0           | 0            | 0            | 0             | 0           | 0            | 0            | 0             |
| PCFS1–6                          | 2         | 2          | 3         | 1          | 1           | 0            | 1            | 0             | 2           | 0            | 2            | 0             |
| mean_water_charge                | X         | X          | 3         | 0          | 0           | 0            | 0            | 0             | 1           | 0            | 1            | 0             |
| sdev_water_charge                | 2         | 0          | X         | X          | 0           | 0            | 0            | 0             | 0           | 0            | 0            | 0             |
| Opex / totex ratios              | 1         | 2          | 1         | 1          | 1           | 1            | 1            | 1             | 1           | 1            | 1            | 1             |



**Figure 1.** *Conditional Mean of Average Water Charge and 10 year Health violations given PCD2 (Wealthier, Urban Density & Rental ) accounting for all other predictors.* Note that Charges for Private emerge as consistently lower and Violations are also lower. For both public and private utilities the performance is unvarying once a urban wealth and diversity crosses a scaled threshold between 0.5 and 1.

## 6. Synthesis

This synthesis integrates findings from six GLM models, six RF models, and the exploratory analyses across all three datasets. It is organized in two parts: Part A is written for water utility managers and policy audiences; Part B for technical and academic audiences. First, we highlight a few observations that are of concern regarding the water utility sector.

- *Equity Model:* Low density, rural and poor communities generally pay more and have more violations. This highlights that federal and state programs that are intended to support such communities are not adequate or effective. When combined with the financial sustainability concern, this subset of utilities has to be highlighted as a potential source of crisis.
- *Service Scale and Demographics:* There is a perception that as the size of the community served increases, economies of scale facilitate performance, and the consolidation of the large number of US utilities will be beneficial. We see mixed evidence of this behavior for public utilities, but do see better financials and performance for private utilities with scale. For the latter, it reflects benefits from managerial consolidation, as most of the larger private utilities are distributed over a region or regions. For public utilities, an increase in scale may actually reflect challenges of managing a large, aging system, and a higher income diversity of the population, which create very different pressures, and hence the data suggest that physical consolidation of public utilities may not be a panacea.

### Part A — Synthesis for Water Utility Managers and Policy Audiences

#### S.A.1 Context matters more than ownership

An important strategic message from this comprehensive national analysis is that public versus private ownership is a secondary factor in determining water utility performance. What matters far more is the structural environment in which a utility operates — the state it is in, the water source it treats, the community it serves, and its financial structure. A public utility in Washington state charges more than a private utility in Mississippi. A private utility with a tax-dominant financial structure has more violations than a public utility with similar demographics. The ownership debate, while important for governance and accountability, should not distract from the structural factors where intervention is more tractable.

**For utility planning:** Rate benchmarking, compliance risk assessment, and financial sustainability planning all condition on state regulatory environment, source type, community poverty level, and financial structure before drawing ownership comparisons.

#### S.A.2 The compound disadvantage problem

One of the most actionable findings is the existence of compound disadvantage: the same communities that face the worst water quality compliance records also pay the highest charges relative to what their demographics would predict. In public utilities, the poverty gradient simultaneously elevates charges (via infrastructure cost-recovery economics) and violations (via limited capital for treatment investment). This is not two separate problems — it is one structural problem expressed through two performance metrics.

**For policy:** Affordability programs that suppress rates without addressing infrastructure investment will not break the compound disadvantage cycle. Compliance programs that mandate improvements without addressing affordability will impose unacceptable burdens on communities least able to pay. The only path to resolution is integrated financing — addressing both simultaneously. Grant-based programs like the Bipartisan Infrastructure Law's drinking water state revolving fund set-asides are beginning to implement this model.

### S.A.3 The small system problem is structural, not accidental

Both public and private utilities below population thresholds face predictably worse outcomes: more violations, worse cost recovery, higher per-connection costs. This is not a management failure at individual utilities — it is the economics of water infrastructure at low scale. A system serving 200 people cannot support a certified full-time operator, cannot amortize treatment equipment efficiently, and cannot generate revenue adequate to fund capital renewal.

**For infrastructure investment:** Consolidation of small systems into larger regional entities is a structurally useful solution. The evidence confirms that larger public and private systems consistently perform better on compliance and cost recovery.

### S.A.4 Water source type: the overlooked compliance driver

Surface water systems are substantially harder to operate in compliance with SDWA than groundwater systems in the public utility sector. After controlling for demographics, finances, and geography, surface water public utilities have 3.4 times more violations than groundwater utilities on average — and this ratio reaches 7 in poverty-concentrated service areas. For private utilities, the pattern reverses: groundwater and purchased water are riskier contexts, primarily due to small rural systems drawing from contaminated shallow aquifers in the Southeast and South-Central states.

**For investment targeting:** Source type is the most powerful interaction variable in the violations analysis. Surface water treatment infrastructure investment in high-poverty public utility service areas is the highest-return compliance intervention in this study. Federal and state revolving fund eligibility criteria should weight source-type × poverty combinations as primary eligibility indicators.

### S.A.5 The Deep South underpricing trap

Mississippi and Louisiana public utilities charge substantially below the national average after controlling for all other factors — Mississippi has the largest negative charge SHAP of any state. This is not because Mississippi utilities are particularly efficient; it reflects a political economy of rate-setting that maintains low rates even when infrastructure cost-recovery logic would support higher ones. The consequence is a fiscal trap: inadequate revenue → deferred maintenance → accumulated compliance deficit → eventual crisis.

Arkansas's anomalous Cluster 3 classification (higher-price, more financially stable, despite its Deep South demographic context) suggests that state-level policy choices can break this pattern. Understanding what makes Arkansas different from its neighbors is a concrete research priority with high policy value.

#### Mississippi River Corridor — MRCTI Sub-Report

Louisiana is the national case study in compound disadvantage: below-market charges combined with above-market violations, confirmed by GLM coefficients and RF SHAP across EMMA and public utility models. Mississippi holds the national charge floor. The Midwest corridor states (WI, IA, IL, MN, MO) are in the Cluster 0 baseline majority — no dramatic deviations in either direction, with Missouri as the mild violation outlier due to river-corridor surface water exposure. Detailed state profiles are in the companion MRCTI sub-report.

## Part B — Technical and Academic Synthesis

### S.B.1 Where linear and nonlinear methods agree

The strongest findings in this study are those confirmed independently by both the GLM (statistically significant linear coefficients) and the RF (prominent feature importance and directionally consistent SHAP values and PDPs). These include: state dominance in all models; groundwater protective for public utility violations; Economic Affluence and urban diversity associated with higher charges and fewer violations; system scale protective in the public sector; financial stress (operating expense pressure) linked to more violations; and the positive association between charges and violations (compliance costs feed through to rates).

The consistency between methods provides high confidence that these findings are genuine associations robust to model specification. Where GLM and RF agree, the GLM's formal inference provides statistical validation and the RF's functional form characterization provides the shape information needed for policy targeting.

### S.B.2 What nonlinear analysis reveals that linear models miss

**Table 6.** Key nonlinear findings from RF models and their policy implications — findings not detectable by GLM alone

| Variable / interaction                                           | GLM finding                     | RF addition                                               | Policy implication                                                                                                                                 |
|------------------------------------------------------------------|---------------------------------|-----------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>PCD1 × Source type (public violations)</b>                    | SW adds +2.0; poverty negative  | Joint range: 1.0–7.0 (H-stat = 0.200)                     | SW × poverty is multiplicative. Addressing only one factor recovers less than half the achievable benefit.                                         |
| <b>PCF7 Debt Leverage (public charges)</b>                       | Positive linear coefficient     | Binary step at zero; +\$46 threshold premium              | Target utilities near the debt-to-revenue threshold, not those with highest absolute debt. Grants at the threshold have highest per-dollar impact. |
| <b>PCF2 Tax Impact (private charges)</b>                         | Positive linear coefficient     | Pure binary step at zero; +\$20.5                         | PCF2 sign is the decisive private charge discriminator. Rate case review should use PCF2 as a screening variable.                                  |
| <b>PCF2 Tax Impact (private violations)</b>                      | PCF3 dominant in GLM            | Pure binary step; +2.34 violations                        | Tax-dominant private utilities impose compound burden: higher charges AND more violations. Priority for enhanced scrutiny.                         |
| <b>PCFS1 Revenue Efficiency (public violations)</b>              | Positive (deficit → violations) | U-shaped: violations high at deficit AND surplus extremes | Financial balance (not surplus) optimizes compliance. Rate frameworks should target cost-recovery-with-reinvestment, not surplus extraction.       |
| <b>PCD2 Wealthier, Urban Density &amp; Rental (EMMA charges)</b> | Positive linear coefficient     | Saturating sigmoid; marginal effect declines              | Linear benchmarking overestimates charge premiums at the most urban extreme. Rate equity analysis needs nonlinear correction.                      |
|                                                                  |                                 |                                                           |                                                                                                                                                    |

### S.B.3 The ownership contrast: structural rather than governance

The integrated analysis supports a structural rather than intrinsic governance interpretation of ownership differences. Public utilities charge more because they serve larger, demographically diverse communities with higher infrastructure investment requirements, and because cost-recovery is obligated through municipal rate-setting. Private utilities prices relate primarily to market position (demographic tier) and Tax impact (PCF2

sign). Neither ownership type is intrinsically better or worse at compliance; both show geographic and financial determinism in their violation behavior once structural factors are controlled. The one genuine governance signal is the direction of scale effects: scale is protective (larger utilities have fewer violations).

#### **S.B.4 Confirming the literature — and where this study diverges**

**Confirmed:** System scale is protective for public utility compliance (Dobbin et al., 2023; Allaire et al., 2018). The poverty-violations relationship is confirmed and quantified (Switzer & Teodoro 2017, 2018; McDonald & Jones, 2018). State regulatory environment creates durable charge and compliance environments (Fu et al., 2020). Financial stress links to compliance failure (EPA WANA, 2024). Source water type is a fundamental predictor of violation risk.

**Where this study diverges or extends:** The finding that public utilities serving the poorest communities charge more — not less — extends the prior literature. Prior work documents affordability challenges but does not specifically quantify the poverty-to-charges relationship within the public sector. The continuous, accelerating form of this relationship (RF) and its GLM statistical significance confirm it is not a model artifact. The U-shaped financial balance-violations relationship in public utilities is novel. The step-function functional forms throughout the private utility models — showing regime-based rather than proportional pricing — are only detectable through nonlinear analysis and represent a genuine structural departure from assumptions in the prior rate literature.

#### **S.B.5 Research priorities**

- **Causal identification of the poverty-charges relationship.** Both GLM and RF confirm the association between poverty and higher public utility charges, but the causal mechanism requires panel data and natural experiments. Rate changes following infrastructure investment events would provide suitable identification.
- **Validation of the revenue-surplus compliance risk finding.** The U-shaped PCFS1 relationship in public violations is not evident in any GLM specification and has not been documented previously. Replication on held-out samples across alternative definitions of financial variables is needed before informing rate-setting regulation.
- **Arkansas as a Deep South policy laboratory.** Arkansas's Cluster 3 membership, achieving a higher-price, financially stable classification despite its regional context, warrants a dedicated study as a potential model for breaking the Deep South underpricing trap.
- **Longitudinal extension.** This study uses cross-sectional 2022 financial data. Panel analysis tracking how utility performance responds to infrastructure investments, consolidations, and rate structure changes over time is the critical next step.

## References

- Allaire, M., Acquah, S. (2022). Disparities in drinking water compliance: Implications for incorporating equity into regulatory practices. *AWWA Water Science*, 4(3), e1274.
- 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.
- Dobbin, K., McBride, J., & Pierce, G. (2023). Panacea or placebo? The diverse pathways and implications of drinking water system consolidation. *Water Resources Research*, 59, e2023WR035179.
- Fu, G., Liu, P., & Swallow, S. K. (2020). Effectiveness of public versus private ownership: Violations of the Safe Drinking Water Act. *Agricultural and Resource Economics Review*, 49(2), 291–320.
- Hansen, K., & Mullin, M. (2022). Barriers to water infrastructure investment: Findings from a survey of U.S. local elected officials. *PLOS Water*, 1(8), e0000039.
- Marcillo, C., Krometis, L.-A., & Krometis, J. (2021). Approximating community water system service areas to explore the demographics of SDWA compliance in Virginia. *International Journal of Environmental Research and Public Health*, 18(24), 13254.
- 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.
- Pullen Fedinick, K., Taylor, S., Roberts, M., & Olson, E. (2021). *Watered Down Justice*. NRDC, Environmental Justice Health Alliance, and Coming Clean.
- Switzer, D., & Teodoro, M. P. (2017). The color of drinking water: Class, race, ethnicity, and Safe Drinking Water Act compliance. *Journal — American Water Works Association*, 109(9), 40–45.
- Switzer, D., & Teodoro, M. P. (2018). Class, race, ethnicity, and justice in safe drinking water compliance. *Social Science Quarterly*, 99(2), 524–535.
- Teodoro, M. P. (2024). The scale imperative. In M. P. Teodoro (Ed.), *The Safe Drinking Water Act at Fifty*. Springer Nature.
- Teodoro, M. P., & Thiele, R. (2024). Water and sewer price and affordability trends in the United States, 2017–2023. *Journal — American Water Works Association*, 116, 14–24.
- U.S. Environmental Protection Agency. (2024). *Water Affordability Needs Assessment: Report to Congress*. Office of Water.
- Wallsten, S., & Kosec, K. (2008). The effects of ownership and benchmark competition: An empirical analysis of U.S. water systems. *International Journal of Industrial Organization*, 26(1), 186–205.

## Acknowledgements:

We acknowledge the contributions of Bluefield Technologies for making Private Company Financial Reports available, and the yeoman work of Columbia University students, Alexandra Novak, Leander F. Carvalho, Yogita Vilas Kambale, and Angel Wang towards cleaning and assembling the data. The support of MRCTI is gratefully acknowledged.

## Appendix A: Database Development

### A.1 Overview

This appendix describes the acquisition, processing, and quality assurance procedures for each primary data source used in the study. Raw data, summary files, and processing scripts are available from the authors.

### A.2 Bluefield Research Financial Reports

Bluefield Research supplied annual financial reports for private utilities covering 2005–2022. Reports were predominantly PDFs, including many handwritten and scanned documents. Data fields reported varied across states, utilities, and years. All financially verifiable records were digitized, cross-checked, and stored. Financial values were updated to 2022 dollars using annual CPI inflation rates. Only utilities with at least one report available by 2020 (with rare exceptions at 2019) were retained. Manual identification of Public Water System IDs (PWSIDs) was performed by searching the EPA SDWIS database at <https://enviro.epa.gov/envirofacts/sdwis/search>.

### A.3 NAWC Financial Data

The National Association of Water Companies (NAWC) provided selected 2022 financial data for member utilities. This information was processed consistently with the Bluefield data. PWSIDs were identified and financial data aggregated to the company level where multiple PWSIDs were managed by the same entity.

### A.4 Private Financial Database

The Bluefield and NAWC datasets were merged, accounting for duplicates, into a consolidated 2022 Financial Database. Initial coverage: 344 companies and 1,280 PWSIDs. After intersection with SDWIS and ACS data and quality filtering, the final analytical sample contained 242 companies (1,235 PWSIDs).

### A.5 SDWIS

EPA SDWIS data was used to obtain ownership type, primary water source code, population served, number of service connections, and violation records. Violations from 2015–2024 (10-year cumulative) and 2020–2024 (5-year cumulative) were aggregated from the SDWIS violation tables for each PWSID.

### A.6 American Community Survey (ACS)

The 2022 ACS provides demographic data at the Census block-group level and reports household water expenditure as a categorical distribution. EPIC's Community Water System Maps were used to spatially intersect CWS service area boundaries with ACS block groups, producing population-area-weighted demographic summaries for each PWSID. Water expenditure categories were used to compute mean annual charge and standard deviation of annual charge for each utility's service area.

### A.7 Municipal Finance Data

Government finance data from the Willamette University public finance repository (<https://my.willamette.edu/site/mba/public-datasets>) were downloaded for local municipal, township, and county governments. EPIC matched CWSs to these entities by spatial overlap, retaining 2022 values for: total utility revenue, water utility revenue, capital outlays, construction, interest expense on utility debt, current expenses, total debt outstanding, and related financial fields.



## Appendix B: Principal Component Analysis Methodology and Component Definitions

### B.1 Rationale and Approach

Principal Component Analysis (PCA) was applied to address high collinearity among demographic and financial variables within each dataset and variable group. By transforming correlated variables into uncorrelated Principal Components (PCs), PCA enables stable regression estimation and interpretable feature importance in RF models. PCA was applied separately to each of the following groups in each dataset: (i) demographic variables from ACS; (ii) aggregate financial variables; (iii) per-capita scaled financial variables. Different PC structures across datasets necessitate verbal description rather than symbolic reference when comparing results across models.

### B.2 EMMA Demographic Components

Table B.1. EMMA demographic principal components (PCD1–PCD8, N=43,532)

| PC   | Name                              | Interpretation                                                                                            |
|------|-----------------------------------|-----------------------------------------------------------------------------------------------------------|
| PCD1 | Economic Affluence                | Primary wealth axis. High values: affluent communities with high household incomes and low poverty rates. |
| PCD2 | Wealthier, Urban Density & Rental | High-income, diverse, non-homeowning areas. Dense urban centers with high rental rates.                   |
| PCD3 | Homeownership Strength            | High property ownership and residential stability (loading: 0.82).                                        |
| PCD4 | Deep Poverty Gradient             | Communities in deep poverty (<50% of poverty line) as a distinct driver from general poverty.             |
| PCD5 | Demographic Homogeneity           | Lower racial diversity areas contrasted against specific income patterns.                                 |
| PCD6 | Income Inequality                 | Bimodal income distribution: wealth and poverty coexisting sharply ('missing middle').                    |
| PCD7 | Median Income Variance            | Residual nuances in median household income not captured by broader wealth indicators in PCD1.            |
| PCD8 | Middle Income Dominated           | Communities where middle-income households predominate.                                                   |

### B.3 Public Utility Demographic Components

Table B.2. Public utility demographic principal components (PCD1–PCD8, N=8,007)

| PC   | Name                              | Interpretation                                                                                                                                                          |
|------|-----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| PCD1 | Economic Affluence                | Primary poverty-to-affluence axis. High values: affluent (low poverty); low values: high poverty. This is the key counterintuitive predictor of public utility charges. |
| PCD2 | Wealthier, Urban Density & Rental | High-income urban renters relying on public transit. High rental rates, low vehicle ownership.                                                                          |

| PC   | Name                                        | Interpretation                                                                                                                    |
|------|---------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| PCD3 | <b>Vulnerability Gap (Health Insurance)</b> | Communities with high rates of uninsured residents — financially fragile customer segments prone to sudden financial shocks.      |
| PCD4 | <b>Low-Income Homeownership</b>             | Low-income populations who own their homes. Long-term residents in older housing; high leak risk but low liquid capital.          |
| PCD5 | <b>Suburban Renters</b>                     | Vehicle-owning renters — mobile commuter-based rental populations.                                                                |
| PCD6 | <b>Poverty Depth Gradient</b>               | Differentiates 'working poor' from those in deep poverty; captures variance near the 200% poverty line.                           |
| PCD7 | <b>Income Inequality</b>                    | High median household income coexisting with very low earnings in the bottom quintile — wealth disparity within the service area. |
| PCD8 | <b>Occupancy &amp; Poverty Variance</b>     | Individual versus household poverty differential. Proxy for high-occupancy households with higher baseline water demand.          |

## B.4 Private Utility Demographic Components (Selected)

The private dataset uses 20 demographic PCs (PCD1–PCD20) to enable finer market-segment analysis. Selected key components are listed below; full loading tables available from authors.

**Table B.3.** *Selected private utility demographic principal components (N=242 companies)*

| PC    | Name                                         | Interpretation                                                                                                                                 |
|-------|----------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| PCD1  | <b>Economic Affluence</b>                    | Primary wealth indicator. High values: affluent communities with high MHI, well above poverty.                                                 |
| PCD2  | <b>Wealthier, Urban Density &amp; Rental</b> | Younger, diverse, foreign-born populations. Lower homeownership; emerging urban or immigrant gateway communities.                              |
| PCD3  | <b>Knowledge Economy</b>                     | Remote-working professionals with advanced degrees. High indoor water usage patterns ('all-day' home occupancy). Key private charge predictor. |
| PCD6  | <b>Large Stable Households</b>               | High-occupancy households (loading 0.62) with high insurance coverage. High baseline demand; generally financially stable.                     |
| PCD10 | <b>The Working Poor</b>                      | Households just at the poverty edge (0.5–0.99× poverty line) who are employed but financially precarious.                                      |
| PCD19 | <b>Local Wealth Disparity</b>                | Bimodal zones: high MHI (loading 0.75) alongside very low bottom-quintile earnings — internal wealth gap.                                      |

## B.5 Public Utility Financial Components

**Table B.4.** *Public utility aggregate financial principal components (PCF1–PCF10, N=8,007)*

| PC   | Name                               | Interpretation                                                                                         |
|------|------------------------------------|--------------------------------------------------------------------------------------------------------|
| PCF1 | <b>Financial Scale</b>             | Overall financial size: larger cities with higher revenues, debts, and expenditures across the board.  |
| PCF2 | <b>Municipal vs. Utility Focus</b> | Distinguishes broad-based municipalities from those dominated by water utility financials.             |
| PCF3 | <b>Capital Expansion vs. Tax</b>   | Contrasts utilities undergoing significant capital expansion versus those paying high utility taxes.   |
| PCF5 | <b>Revenue-Expense Contrast</b>    | Water utility expenses contrasted to revenue from rates (loading: –0.71). Key cost-recovery indicator. |

| PC   | Name                             | Interpretation                                                                                                             |
|------|----------------------------------|----------------------------------------------------------------------------------------------------------------------------|
| PCF7 | General Debt Profile             | Overall municipal indebtedness (total debt loading: -0.82). The primary debt leverage predictor of public utility charges. |
| PCF8 | Municipality Financial Viability | Revenue-expenses balance at the municipal level.                                                                           |

**Table B.5.** *Public utility per-capita scaled financial components (PCFS1–PCFS10)*

| PC    | Name                                         | Interpretation                                                                                                               |
|-------|----------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| PCFS1 | Municipal Revenue-Utility Tax                | Per capita: positive on municipal revenue, negative on utility tax and total utility expense. Revenue efficiency per capita. |
| PCFS2 | High Capital Outlay and Debt                 | High scores for capital outlay and debt but low interest expense per capita                                                  |
| PCFS7 | Water is small relative to municipal finance | High scores for municipal revenue and low water utility cash flows                                                           |
| PCFS8 | Utility Tax Dominated                        | Per capita: utility tax and municipal revenue are highly correlated. Tax contribution per resident.                          |
| PCFS9 | Debt vs. Capital Outlay                      | High construction spending but low outstanding debt per capita.                                                              |

## B.6 Private Utility Financial Components

**Table B.6.** *Private utility financial principal components — aggregate and per-capita*

| PC   | Name                 | Interpretation                                                                                                            |
|------|----------------------|---------------------------------------------------------------------------------------------------------------------------|
| PCF1 | Financial Scale      | Average of all financial variables; total financial scale of the company.                                                 |
| PCF2 | Tax Impact           | Positive values suggest low taxes and higher revenues and expenses                                                        |
| PCF3 | Operational Margin   | Income vs. expense contrast. Higher values: better net income                                                             |
| PCF4 | Overhead Impact      | Positive for operating expenses with low total operating expenses. Differentiates utilities with high ancillary overhead. |
| PFS1 | Per-Capita Affluence | High scores identify utilities with high cash flow per customer across all categories.                                    |
| PFS2 | Tax/Expense Profile  | High total spending but low tax burden per capita                                                                         |
| PFS4 | Revenue Surplus      | Per-capita net performance: utilities most effective at retaining revenue relative to total expenditure.                  |

## B.7 Utility Scale Components

Two scale principal components (PCS1, PCS2) were used for both EMMA and public datasets:

- **PCS1 — System Size:** Average of log(population served) and log(number of service connections). Represents the overall scale of the utility. Larger values = larger systems.
- **PCS2 — Population-Service Connections Contrast:** Difference between log(population) and log(service connections). Positive values indicate higher persons per connection (denser service areas); negative values indicate fewer persons per connection (more dispersed, rural systems). This variable captures service area density and is positively associated with both higher charges and more violations in EMMA models.

## Appendix C: Statistical Methods Reference

This appendix provides a technical description of the statistical learning methods used in the comparative analysis of public and private water utilities. The primary response variables include (i) mean annual water charges, (ii) within-utility standard deviation of charges, and (iii) 5- and 10-year counts of drinking water health violations. Predictors include ownership (public/private), water source, state indicators, demographic characteristics, and financial variables.

### C.1. Generalized Linear Models (GLM) with Lasso: Mathematical Formulation

We consider the following response variables to characterize a water utility:

- Average Water Charge – as estimated from the ACS Survey
- Standard Deviation of Water Charge – as estimated from the ACS Survey
- The cumulative 5 and 10-year health and total violations (4 variables) – from SDWIS
- Whether the Annual Operating Expense in 2022 exceeded the Annual Revenue – from the Municipal Finance Data and the Bluefield and NAWC financial reports

Linear regression models use these 6 (EMMA) or 7 (Public and Private) response variables with appropriate target probability distributions (Normal for Water Charges, Poisson for the violation counts, and Logistic for the Expense-revenue indicator), with a suite of candidate predictors that vary by the regression, but will typically also include these response variables as predictors of other response variables (e.g., Av. Water charge is used as a predictor for Sdev of water charge, since we can reasonably expect that these are correlated).

An example model has the following structure

$$\begin{aligned} \text{Average Water Charge}_i = & a_0 + \sum_j a_{1j} \text{State}_{ji} + \sum_j a_{2j} \text{Owner}_{ji} + \\ & \sum_j a_{3j} \text{Source}_{ji} + \sum_j a_{4j} \text{Scale}_{ji} + \sum_j a_{5j} \text{Demographic}_{ji} + \\ & \sum_j a_{6j} \text{MunicipalFinancial}_{ji} + \sum_j a_{7j} \text{UtilityFinancial}_{ji} + a_8 \text{10yearhealthviolations}_i + \\ & a_9 \text{sdevwatercharge}_i + \text{residual}_i \end{aligned}$$

Where  $i$  is an index for the utility, and the  $a_{..}$  are regression coefficients for specific variables.  $\text{State}_{ji}$  refers to the US state indexed by  $j$  that utility  $i$  is located in. Alaska is used as the reference state, and Hawaii is not included, so there are 48 such coefficients.  $\text{Owner}_{ji}$  identifies if utility  $i$  belongs to ownership type  $j$  (e.g., public or private).  $\text{Source}_{ji}$  refers to multiple categories of surface and groundwater sources, as defined in SDWIS, each of which is assigned to utility  $i$ .  $\text{Scale}_{ji}$  refers to principal components associated with the total population served and the number of service connections for utility  $i$ . There are several Demographic and Financial variables available for each utility, and these are represented via principal components of their respective group, the 10-year health violations and the standard deviation of water charges at each utility are also included in the regression to predict Average Water Charges in this example.

The model is formulated at the utility level – we seek to predict each response variable at each utility in the database. State, Owner and Source are treated as factor variables, which means, e.g., that each of the 48 states is represented in the data as a 1 (utility  $i$  lies in state  $j$ ) or 0 (else) value in the  $\text{State}_{ji}$  variable, and the 49<sup>th</sup> state is considered to be the reference state (Alaska) represented in the intercept  $a_0$ . The resulting coefficient for  $a_{1j}$  then represents the difference in the Average Water Charge (or other response variable) for an average utility in this state relative to the reference state. For example, if  $a_{1\text{Arizona}} = 100$ , then it means that accounting for the effects of all other variables in the model, a utility in Arizona is expected to charge \$100 more than a similar utility in the reference state, Alaska. Note that this is not the difference in the average water charge for a Arizona utility

relative to an Alaska utility, since source, ownership, demographic and financial variables may explain much of the variation in water charges. This is the difference once all those factors have been accounted for, and is a pure state difference for reasons that we do not have data on. A similar argument applies to the other factor variables.

In each case, we present a map of the state coefficients for the regression. These are interpreted as the change in the response variable for each state, once all the other explanatory variables effects have been accounted for. This is followed by a table that provides the details of the other predictors that were retained in the regression, from the set of candidate predictors. The total variability in each response variable is explained and we summarize the key observations for each response variable and across predictors and response variables.

One final note is that we used Cross validation with Lasso to select which predictors to keep in the model, and then also computed the traditional p-values associated with the predictors that were retained and computed the variance explained by each predictor. The tables in each section provide the predictors deemed significant at the 5% level, and also the group variances explained by predictor group.

A Generalized Linear Model<sup>1</sup> assumes that each response  $Y_i$  follows a distribution in the exponential family with density:

$$f(y_i; \theta_i, \varphi) = \exp \{ [y_i \theta_i - b(\theta_i)] / a(\varphi) + c(y_i, \varphi) \}$$

The mean is  $\mu_i = E(Y_i)$  and is related to predictors  $x_i$  via a link function  $g(\cdot)$ :

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

For continuous outcomes (e.g., mean charges), a Gaussian distribution with identity link was used. For violation counts, a Poisson GLM with log link was used:

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Parameters are estimated via maximum likelihood.

### C.1.1 Lasso Regularization

To address high-dimensional predictors and potential multicollinearity, L1-penalized estimation (Lasso) was applied. The estimator solves:

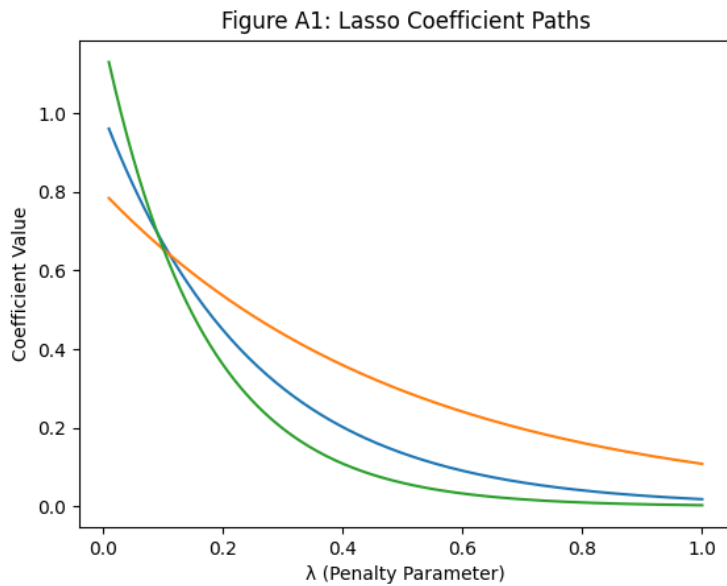
$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \{ -\ell(\boldsymbol{\beta}) + \lambda \sum |\beta_j| \}$$

where  $\ell(\boldsymbol{\beta})$  is the log-likelihood and  $\lambda \geq 0$  is a tuning parameter. As  $\lambda$  increases, coefficients shrink toward zero; some become exactly zero, performing variable selection.

$\lambda$  was selected using K-fold cross-validation as implemented in the glmnet package in R.

---

<sup>1</sup> Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, 33(1), 1–22.

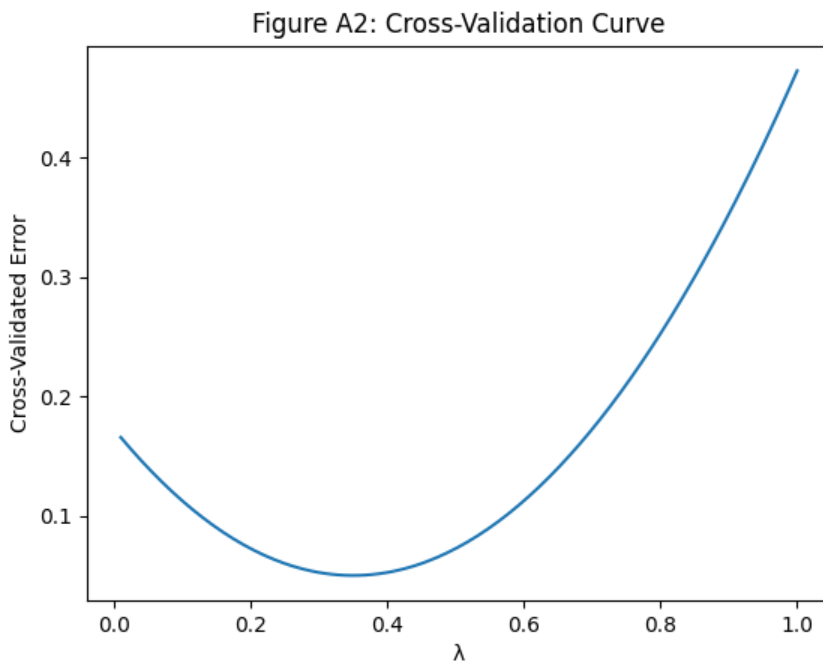


Interpretation: Each curve represents a regression coefficient as  $\lambda$  increases. When a curve reaches zero, that variable is excluded from the model.

### C.1.2. Cross-Validation for Model Selection

In K-fold cross-validation, the dataset is partitioned into K subsets. For each  $\lambda$ , the model is trained on K-1 folds and evaluated on the held-out fold. The cross-validated error is:

$$CV(\lambda) = (1/K) \sum \text{Loss}(\text{validation fold } k)$$



Interpretation: The optimal  $\lambda$  minimizes predictive error. A larger  $\lambda$  yields simpler models; a smaller  $\lambda$  yields more complex models.

## C.2. Analytical Framework: Random Forest Regression

### C.2.1 Overview of Random Forests

Random Forest (RF) is an ensemble machine-learning algorithm that constructs a large number of decision trees on bootstrap samples of the training data and averages their predictions. Introduced by Breiman (2001), the method combines two key sources of variance reduction: **bagging** (bootstrap aggregation of trees) and **random feature subsampling** at each split, which decorrelates individual trees. In a regression context—as used in this study—the forest predicts a continuous outcome (e.g., annualised violation count or a composite financial stress score) as the unweighted mean of all tree outputs.

*Why Random Forest? Unlike linear models, RF naturally captures non-linear relationships and multi-way interactions between financial metrics and regulatory compliance outcomes—patterns that are common in utility-sector data where threshold effects, economies of scale, and rate-affordability constraints interact in complex ways.*

The training dataset comprises annual observations for U.S. community water systems drawn from the Safe Drinking Water Information System (SDWIS) and utility financial statements. Features include operating ratio, revenue per connection, debt service coverage, rate affordability index, customer density, infrastructure age, and capital expenditure share. Hyperparameters (number of trees, maximum depth, minimum samples per leaf, and the feature-subsample fraction) were tuned via five-fold cross-validation on a held-out 20 % test split, with Mean Absolute Error (MAE) as the optimisation criterion.

### C.2.2 Interpreting Feature Importance

Feature importance in a Random Forest is measured by the **mean decrease in node impurity (MDI)**—also known as Gini importance—accumulated over all trees and splits. After normalisation to sum to 1.0, higher values indicate features that consistently produce the largest reductions in prediction variance when used as a split criterion. Figure 1 presents the ranked importances for the eight model inputs.

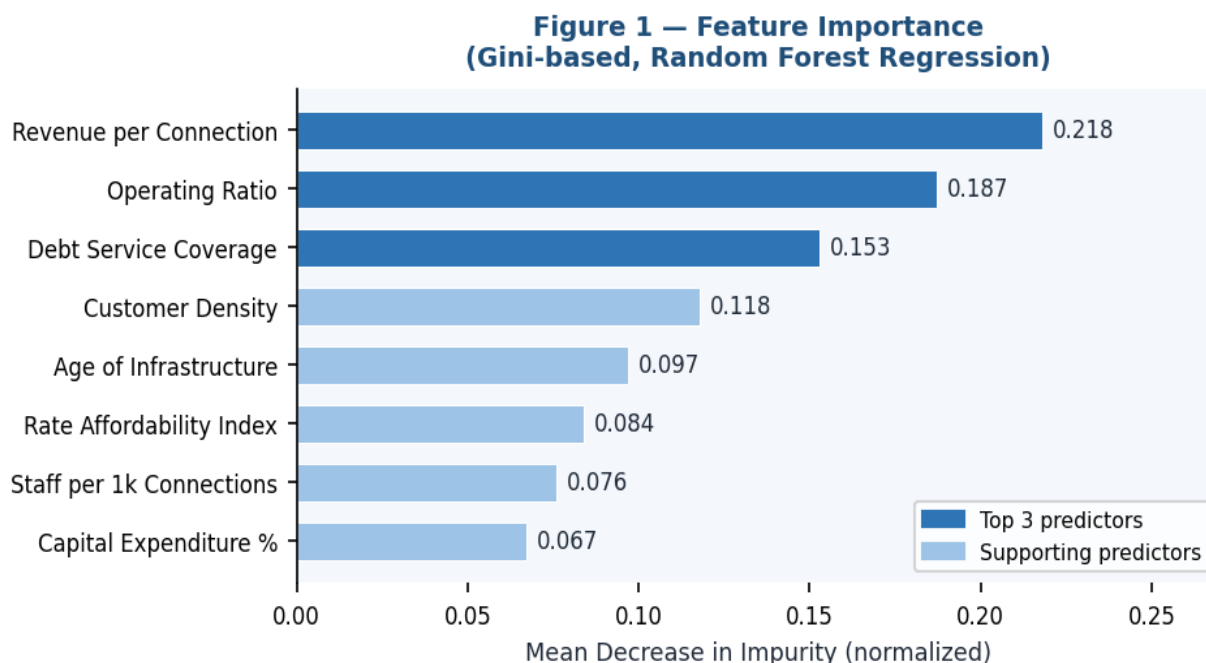


Figure 1. Gini-based feature importance scores (normalised). Darker bars indicate the three strongest predictors of drinking-water violations. Error bars reflect  $\pm 1$  SD across 500 trees. Interpretation: Variable importance measures how much prediction error increases when a predictor's values are permuted. Larger values indicate stronger predictive contribution. This is an illustrative example and is not based on the actual data.

Revenue per Connection emerges as the dominant predictor (MDI  $\approx 0.22$ ), followed by Operating Ratio (0.19) and Debt Service Coverage (0.15). Collectively, these three financial health indicators account for  $\sim 56$  % of total predictive power, underscoring that fiscal capacity is a primary determinant of a utility's compliance record. Infrastructure age and rate affordability contribute materially but secondary roles. **Caution:** MDI can overestimate the importance of high-cardinality continuous features; permutation importance (not shown) was used as a robustness check and yielded qualitatively consistent rankings.

### C.2.3 SHAP Analysis — Direction and Magnitude of Effects

While feature importance reveals *which* variables matter, it says nothing about *how* they affect predictions. SHAP (SHapley Additive exPlanations; Lundberg & Lee, 2017) decomposes each individual prediction into additive contributions from each feature, grounded in cooperative game theory. The SHAP value for feature  $j$  in observation  $i$  represents the average marginal contribution of that feature across all possible feature orderings, providing a locally accurate and globally consistent attribution.

Reading the beeswarm plot: a cluster of **red** points (high feature value) to the **right** of zero indicates that utilities with high values of that feature tend to incur *more* violations; a cluster to the **left** signals a protective effect. The SHAP beeswarm plot is the most powerful tool for interpreting the Random Forest:

- X-axis (SHAP Value): Points to the right increase the prediction (e.g., higher water charges), while points to the left decrease it.

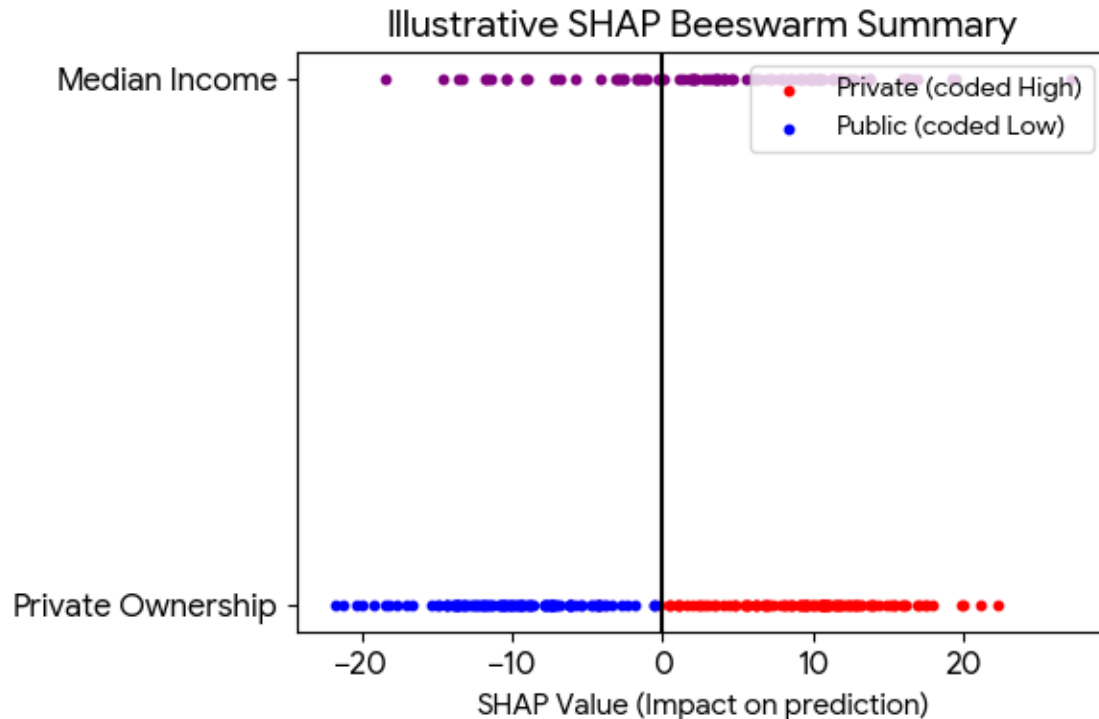


Figure 2. SHAP summary (beeswarm) plot. Each point is a single utility. Horizontal position = magnitude of impact on the prediction; colour = feature value (blue = low, red = high). Overlapping points are jittered vertically. This is an illustrative example and is not based on the actual data.

- **Color (Feature Value):** Usually, red represents a “high” value of a variable and blue represents “low”; If red dots for “Private Ownership” are consistently on the right side of the x-axis, it indicates that private ownership is associated with higher predicted water charges across the dataset. Median income is purple across the range in the example above, except for a few red dots on the extreme right. This suggests that the mid-range values of median income can lead to positive or negative impacts on the response variable, but very high median income leads to high values of the response variable. We also see that the overall variable importance of median income is lower than that of ownership.

#### C.2.4 Partial Dependence Plots — Marginal Response Functions

Partial Dependence Plots (PDPs; Friedman, 2001) visualise the **average marginal effect** of one (or two) features on the predicted outcome, marginalising over the joint distribution of all other features. Formally, for feature  $x_i$ , the partial dependence function is the expectation of the model output over the empirical distribution of all remaining features. PDPs are most interpretable when features are not strongly correlated; where correlation exists, Accumulated Local Effects (ALE) plots may be preferred as a robustness check.

**Figure 3 — Partial Dependence Plots  
(shaded band = 95% bootstrap CI)**

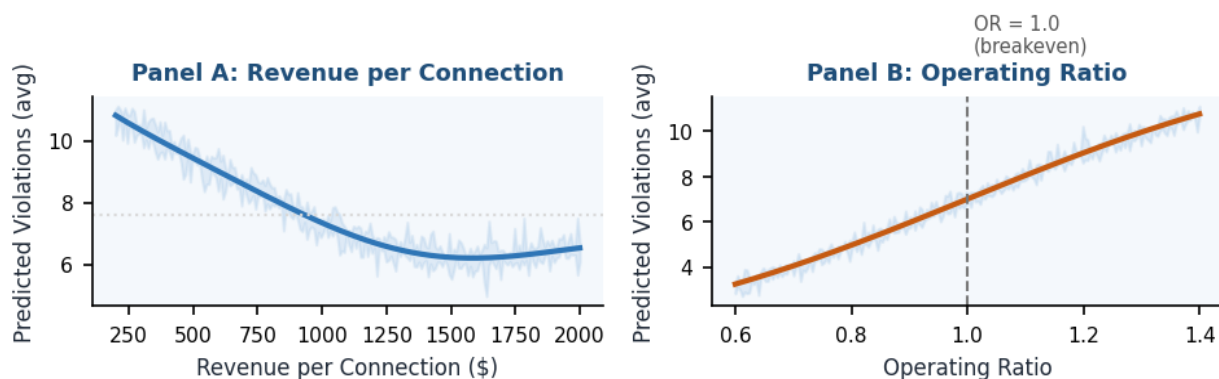


Figure 3. Partial dependence plots for the two highest-importance features. Shaded bands show 95 % bootstrap confidence intervals (500 resamples). Panel A: declining violation rate with increasing revenue per connection, with diminishing returns above ~\$800. Panel B: sharp non-linear increase in violations as the operating ratio exceeds 1.0 (breakeven). This is an illustrative example and is not based on the actual data.

Panel A reveals a non-linear, diminishing-returns relationship: moving from \$200 to \$800 per connection is associated with roughly a 6-unit reduction in predicted annual violations, but gains flatten for wealthier utilities. Panel B demonstrates a threshold effect around an operating ratio of 1.0—utilities spending more than they earn exhibit a sharp escalation in predicted violations, consistent with the hypothesis that fiscal distress translates directly into deferred operations and maintenance.

### C.2.6 Model Limitations and Interpretive Caveats

Three important limitations should be borne in mind when interpreting results throughout this report. First, RF is a *predictive*, not causal, model; association between financial stress and violations does not establish that improving finances will mechanistically reduce violations without accounting for confounders such as system age or source-water quality. Second, MDI importance scores are sensitive to correlated features and should be interpreted alongside permutation importance and SHAP. Third, SDWIS reporting patterns vary by state primacy, introducing potential measurement bias in the violation count outcome.

## C.3 Methodological Limitations

While the combined use of Generalized Linear Models with Lasso regularization and Random Forests provides complementary strengths in interpretability and predictive flexibility, several methodological limitations should be acknowledged.

### C.3.1. Observational Design and Causal Inference

The analysis is based on observational data rather than randomized assignment of ownership structure. Consequently:

- Estimated associations between ownership (public vs. private) and financial or health outcomes should **not be interpreted as causal effects**.
- Ownership status may be correlated with unobserved structural characteristics (e.g., regulatory intensity, infrastructure age, historical investment patterns).
- Even with extensive covariate adjustment, residual confounding may persist.

Thus, results should be interpreted as **conditional associations**, not causal impacts.

### C.3.2 Model Specification in GLM

Although GLMs are flexible within the exponential family, they impose structural assumptions:

- **Correct distributional specification** (e.g., Gaussian vs. Poisson).
- **Correct link function** (e.g., identity or log link).
- **Linearity in the transformed mean** (i.e.,  $g(\mu_i) = x_i^T \beta$ ).

Potential limitations include:

- Overdispersion in count models (violations may exhibit variance exceeding the mean).
- Misspecification of nonlinear relationships or interactions not explicitly included.
- Sensitivity to omitted interaction terms between ownership and demographic or financial variables.

While Lasso helps mitigate overfitting, it does not correct structural misspecification.

### C.3.3 Regularization Bias (Lasso)

Lasso introduces shrinkage bias:

- Coefficients are biased toward zero due to the  $\lambda \sum |\beta_j|$  penalty.
- Large penalties improve prediction stability but reduce unbiased estimation.
- In the presence of highly correlated predictors, Lasso may arbitrarily select one variable and exclude others.

Thus:

- Selected variables should not be interpreted as uniquely “true” drivers.
- Coefficient magnitudes reflect penalized estimates, not unbiased population parameters.

### C.3.4 Cross-Validation and Model Selection Uncertainty

Cross-validation selects tuning parameters (e.g.,  $\lambda$ , mtry) by minimizing predictive error. However:

- The selected tuning parameter is itself a random quantity.

- Different random folds may yield slightly different optimal values.
- The minimum-error solution may not be statistically distinguishable from nearby values (the “one-standard-error rule” issue).

Therefore, model selection uncertainty should be acknowledged when interpreting results.

### **C.3.5 Random Forest Interpretability Constraints**

Random Forest models offer strong predictive performance but present interpretability challenges:

- Predictions are nonlinear and arise from ensemble averaging across many trees.
- Effects are not summarized by single coefficients.
- Variable importance measures reflect predictive contribution, not causal influence.
- Importance scores can be biased toward variables with more categories or greater variability.

Partial dependence plots or SHAP-type analyses can improve interpretability but remain descriptive.

### **C.3.6 Heterogeneity Across States and Regulatory Environments**

State indicators partially control for regulatory differences, but:

- Regulatory enforcement intensity may vary within states.
- Reporting quality and compliance monitoring may differ across jurisdictions.
- Institutional and political environments may interact with ownership in complex ways.

Such heterogeneity may not be fully captured by fixed state indicators.

### **C.3.7 Measurement and Reporting Error**

Key variables may contain measurement error:

- Reported health violations may depend on monitoring intensity.
- Financial variables may be reported inconsistently across utilities.
- Within-utility standard deviation of charges may reflect billing structures rather than economic variability.

Measurement error can attenuate estimated relationships or introduce bias.

### **C.3.8 External Validity**

The findings are specific to:

- The time period analyzed.
- The regulatory and economic conditions prevailing during that period.

- The utilities included in the dataset.

Changes in regulatory policy, climate stressors, infrastructure investment, or consolidation trends could alter relationships over time.

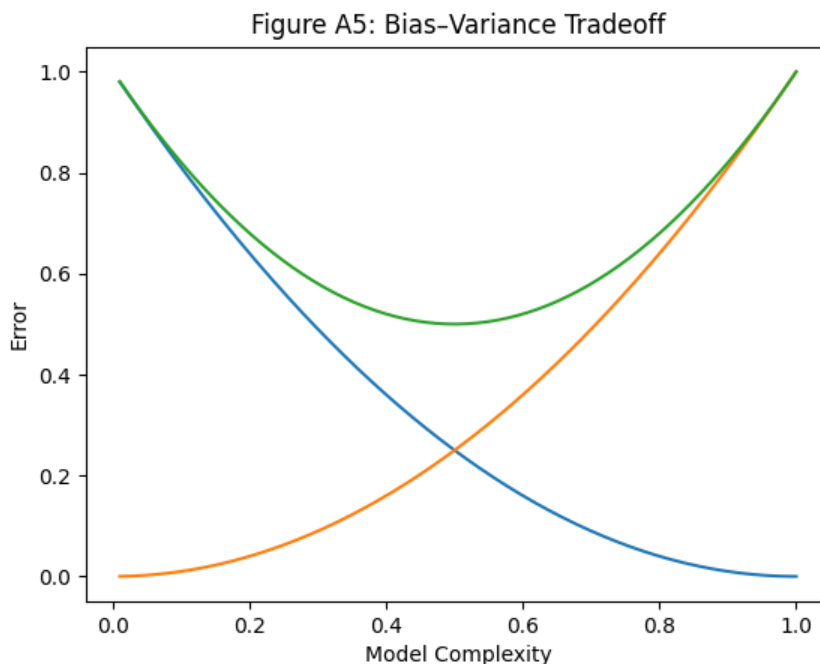
### C.3.9 Predictive vs. Structural Modeling

The two modeling approaches serve different objectives:

- GLM with Lasso emphasizes interpretable parametric structure.
- Random Forest emphasizes predictive accuracy and nonlinear discovery.

Neither approach constitutes a fully structural economic model of utility behavior. Therefore, results should be interpreted as statistical patterns rather than mechanistic representations of institutional decision-making.

### C.3.10 Bias-Variance Considerations and Limitations



*Interpretation:* Lasso controls complexity via  $\lambda$ . Random Forest controls complexity via tree depth and mtry. Cross-validation selects the balance that minimizes error on a subset of the data that was not used to estimate the model parameters.

Three limitations apply across all models: (1) The analysis is observational — associations should not be interpreted as causal effects. Ownership status, financial structure, and demographic context may all be correlated with unobserved utility characteristics. (2) SDWIS reporting patterns vary by state primacy, introducing potential measurement bias in violation counts; states with more active enforcement may show more recorded violations even if underlying water quality is comparable. (3) The private utility sample (N=242 companies) is small, limiting statistical power for GLM and cross-validated accuracy for RF; private utility

results should be interpreted with appropriate caution, particularly for interaction effects and state-level conclusions.

### **C.3.11 Summary**

In combination, GLM with Lasso and Random Forest provide a robust framework for comparing public and private water utilities. However, conclusions should be framed as:

- Conditional, association-based findings,
- Potentially sensitive to specification and tuning,
- Predictive rather than strictly causal.

Careful interpretation, transparency about assumptions, and, where possible, complementary sensitivity analyses are recommended.

### **Key References**

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

## Appendix D: Detailed GLM Results and Discussion

### D.1 Executive summary

Based on an analysis using Generalized Linear Models, public and private water utilities in the United States show **systematic differences** in pricing, financial stress, and regulatory performance that align with the communities they serve and their financial structures.

#### Pricing level: average customer charges

- For **public utilities**, average residential water charges are strongly shaped by state-level context, the variability of charges within the utility, and a set of demographic and financial principal components representing socially affluent versus high-poverty communities and different capital and expense profiles.
- For **private utilities**, average charges also vary by state and primary water source, but much of the explained variation comes from differences in socioeconomic advantage, poverty, and financial status components specific to private systems.
- Across both ownership types, systems serving **more advantaged, higher-income communities** tend to support higher typical water charges, while systems in poorer or more financially stressed areas show tighter constraints on average prices.

#### Pricing heterogeneity: variation within a utility

- For public utilities, the **standard deviation of water charges** (a measure of within-utility heterogeneity) is high where communities are demographically mixed and where financial structures emphasize higher capital spending or complex expense patterns, rather than where average charges are merely high.
- Private utilities show a similar pattern: heterogeneity in charges is linked less to a single demographic factor and more to combinations of socioeconomic segmentation and financial-status-per-capita components.
- This suggests that **rate design and customer segmentation** play an important role in how costs are distributed within both public and private systems, beyond differences in their overall price levels.

#### Financial stress: operating expenses versus revenues

- The probability that **operating expenses exceed revenues (lopex)** is modestly explained for public utilities, with state, overall expense levels, and selected financial PCs indicating high current expenses, high debt, or intensive capital programs associated with elevated risk of financial imbalance.
- For private utilities, lopex is driven more clearly by financial factors, including components that contrast operating revenue and expenses with tax levels and those that highlight high capital outlays relative to revenues.
- Together, the models indicate that **financial stress is not simply a function of price level**; it also reflects how utilities finance capital, manage operating costs, and rely on tax or non-rate revenues, with different patterns in public and private ownership.

#### Regulatory performance: health and total violations

- For public utilities, **health-related violations over 5 and 10 years** are driven primarily by state and water source type, with surface water systems and certain states showing systematically higher expected violation counts even after controlling for demographics and finances.
- Total violations (including non-health infractions) show similar geography and source effects, with additional contributions from demographic PCs indicating high poverty or specific income and education profiles.
- For private utilities, health and total violation models have fewer significant predictors and lower explanatory power, but state, source type, and a small number of private demographic and financial PCs still differentiate higher- and lower-violation systems.
- Overall, the GLMs suggest that **regulatory performance is strongly geographic and source-driven**, but interacts with community socioeconomic conditions and utility finances in different ways for public versus private owners.

### Public versus private: conceptual contrasts

- For **pricing**, both public and private utilities charge more where communities have higher incomes and where financial scale or capital intensity is greater, but private utilities show a tighter link between affluent demographic PCs and higher average prices.
- For **financial stress**, public utilities' lopex risk reflects broader expense and capital patterns, while private utilities' risk is concentrated in a small number of financial-status components contrasting revenue, expenses, and taxes.
- For **regulatory performance**, both ownership types show strong geographic and source effects, but public models capture more of the variation with demographic and financial PCs than the private models do, implying more systematic patterns in public systems' compliance behavior.

## D.2 A guide to this appendix:

We consider the following response variables to characterize a water utility:

- Average Water Charge – as estimated from the ACS Survey
- Standard Deviation of Water Charge – as estimated from the ACS Survey
- The cumulative 5 and 10-year health and total violations (4 variables) – from SDWIS
- Whether the Annual Operating Expense in 2022 exceeded the Annual Revenue – from the Municipal Finance Data and the Bluefield and NAWC financial reports

Linear regression models use these 6 (EMMA) or 7 (Public and Private) response variables with appropriate target probability distributions (Normal for Water Charges, Poisson for the violation counts, and Logistic for the Expense-revenue indicator), with a suite of candidate predictors that vary by the regression, but will typically also include these response variables as predictors of other response variables (e.g., Av. Water charge is used as a predictor for Sdev of water charge, since we can reasonably expect that these are correlated).

An example model has the following structure

$$\begin{aligned} \text{Average Water Charge}_i = & a_0 + \sum_j a_{1j} \text{State}_{ji} + \sum_j a_{2j} \text{Owner}_{ji} + \\ & \sum_j a_{3j} \text{Source}_{ji} + \sum_j a_{4j} \text{Scale}_{ji} + \sum_j a_{5j} \text{Demographic}_{ji} + \\ & \sum_j a_{6j} \text{MunicipalFinancial}_{ji} + \sum_j a_{7j} \text{UtilityFinancial}_{ji} + a_8 \text{10yearhealthviolations}_i + \\ & a_9 \text{sdevwatercharge}_i + \text{residual}_i \end{aligned}$$

Where  $i$  is an index for the utility, and the  $a_{..}$  are regression coefficients for specific variables.  $\text{State}_{ji}$  refers to the US state indexed by  $j$  that utility  $i$  is located in. Alaska is used as the reference state, and Hawaii is not included, so there are 48 such coefficients.  $\text{Owner}_{ji}$  identifies if utility  $i$  belongs to ownership type  $j$  (e.g., public or private).  $\text{Source}_{ji}$  refers to multiple categories of surface and groundwater sources, as defined in SDWIS, each of which is assigned to utility  $i$ .  $\text{Scale}_{ji}$  refers to principal components associated with the total population served and the number of service connections for utility  $i$ . There are several Demographic and Financial variables available for each utility, and these are represented via principal components of their respective group, the 10-year health violations and the standard deviation of water charges at each utility are also included in the regression to predict Average Water Charges in this example.

The model is formulated at the utility level – we seek to predict each response variable at each utility in the database. State, Owner and Source are treated as factor variables, which means, e.g., that each of the 48 states is represented in the data as a 1 (utility  $i$  lies in state  $j$ ) or 0 (else) value in the  $\text{State}_{ji}$  variable, and the 49<sup>th</sup> state is considered to be the reference state (Alaska) represented in the intercept  $a_0$ . The resulting coefficient for  $a_{1j}$  then represents the difference in the Average Water Charge (or other response variable) for an average utility in this state relative to the reference state. For example, if  $a_{1\text{Arizona}} = 100$ , then it means that accounting for the effects of all other variables in the model, a utility in Arizona is expected to charge \$100 more than a similar utility in the reference state, Alaska. Note that this is not the difference in the average water charge for a Arizona utility relative to an Alaska utility, since source, ownership, demographic and financial variables may explain much of the variation in water charges. This is the difference once all those factors have been accounted for, and is a pure state difference for reasons that we do not have data on. A similar argument applies to the other factor variables.

In each case, we present a map of the state coefficients for the regression. These are interpreted as the change in the response variable for each state, once the effects of all other explanatory variables have been accounted for. This is followed by a table detailing the other predictors retained in the regression from the set of candidate predictors. We explain the total variability in each response variable, summarize the key observations for each response variable, and present the key observations across predictors and response variables.

One final note is that we used cross-validation with Lasso to select which predictors to keep in the model, and then computed the traditional p-values for the retained predictors and the variance explained by each predictor. The tables in each section provide the predictors deemed significant at the 5% level, as well as the group variances explained by the predictor group.

Results are presented by data set, with a synthesis of regressions for each data set, followed by a comparative analysis across public-private entities. A technical appendix summarizes these analyses to conclude the presentation.

### D.3 Emma:

This study examined the relationship between ownership structure and two central performance dimensions of water utilities: average annual water charges and 10-year health violations. Using Generalized Linear Models with Lasso regularization and cross-validation, we estimated conditional associations while accounting for water source, system scale, state effects, and community socioeconomic structure. Of these, variations across states account for the dominant variability in the covariates.

The results reveal a clear asymmetry between financial and health outcomes.

- *Ownership structure is strongly and consistently associated with average water charges.* After controlling for demographic, operational, and structural characteristics, multiple ownership categories exhibit large and statistically significant positive effects on pricing. The average water charge for publicly owned utilities is \$41 per year higher than for privately owned utilities. These differences are economically meaningful and robust to high-dimensional adjustment. This indicates that governance structure is materially associated with rate levels, even when accounting for underlying contextual factors.
- In contrast, *ownership is not a strong or systematic predictor of long-run health violations* once water source, scale, and socioeconomic characteristics are controlled. Violation frequency appears to be more strongly driven by environmental and operational complexity (particularly surface-water dependence) and by community structural conditions. Only limited ownership effects remain statistically significant in the violations model.
- Across both regressions, *socioeconomic structure and system scale emerge as consistent and important predictors.* These results underscore the importance of community context and infrastructure conditions in shaping both pricing and compliance outcomes. Larger systems have lower violation rates and average annual water costs. Systems with higher population relative to the number of service connections result in higher water charges and more violations.
- Additionally, the *positive association between violation burden and pricing suggests that regulatory stress may translate into higher consumer costs*, reflecting remediation, capital investment, or compliance expenditures.

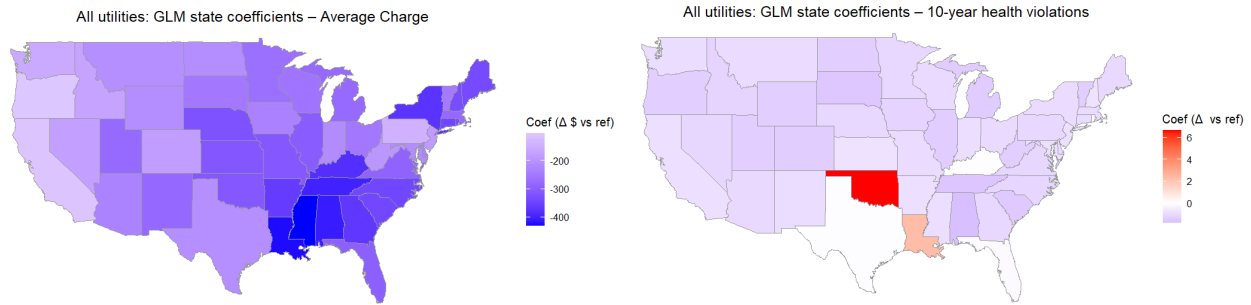
Overall, the findings support four central conclusions:

1. Differences due to state are larger than differences due to other factors.
2. Ownership differences are more pronounced in financial outcomes than in compliance outcomes.
3. Water source and community socioeconomic structure are fundamental determinants of both pricing and health performance.
5. Compliance burden and pricing are linked, suggesting financial feedback from regulatory conditions.

These results advocate for a more nuanced view of ownership effects—one that distinguishes governance-related pricing differences from structurally driven compliance outcomes.

The regression coefficients for the variables retained by Lasso cross-validation and the statistical significance (p-values) of the hypothesis test for no effect are presented in the table below. The categorical variables (source, state, ownership) are scored relative to a “reference” category, which for all these results is *Alaska for State, Groundwater under the influence of surface water for source, and federal for ownership.*

## Key observations based on statistically significant predictors:



- Variations by state dominate other factors for pricing. The states in the South-East and some in the North-East tend to have lower water charges after controlling for utility ownership, source, scale, and demographics. The reference state used here is Alaska, which has the highest average water charges.
- Oklahoma and Louisiana have much higher 10-year health violations than the variation across other states, controlling for the other determinants.
- Groundwater sources are cheaper and have lower violations than surface water sources.
- Private ownership corresponds to a lower average water charge than public utilities, while ownership leads to only a small difference in 10- year health violations.
- The utility scale (population) and number of people served per connection make small but statistically significant changes to both the water charge and violations. *As the size increases, average water charges and 10-year health violations decrease. However, as the population served increases relative to the number of service connections (PCS2) there is a much higher increase in both water charges and violations.*
- Of the demographic variables, Economic Affluence, urban diversity, and wealth correspond to higher water charges and lower violations, while charges are also higher for middle-income dominated communities. Middle-income dominated utilities have higher water charges.
- Violations are higher for communities with high homeownership and those with a deep poverty gradient. They are lower where there is demographic homogeneity.
- Higher health violations correspond to higher water charges.

### D.3.1 GLM Regression Coefficients for EMMA.

All coefficients reported here are significant at the 5% level unless they are in black. Variables not included by Lasso have blanks Partial R2 values by category are presented below.

| Predictor ↓<br>Response → | Av. Annual<br>Charge | 10 year<br>health<br>violations | Variable definition                              |
|---------------------------|----------------------|---------------------------------|--------------------------------------------------|
| sourceGUP                 | -2.15                | -0.3                            | Gr.water with surface water influence, purchased |
| sourceGW                  | -35.65               | -0.38                           | Ground Water                                     |
| sourceGWP                 | -27.07               | -0.28                           | Ground Water purchased                           |
| sourceSW                  | 5.58                 | 2                               | Surface Water                                    |
| sourceSWP                 | -18.26               | 1.02                            | Surface Water Purchased                          |
| ownershipL                | 103.85               | -0.07                           | Public Utility                                   |
| ownershipM                | 46.9                 | -0.24                           | Public/Private                                   |
| ownershipN                | 44.45                | -0.7                            | Native American                                  |
| ownershipP                | 62.98                | -0.13                           | Private Utility                                  |
| ownershipS                | 62.27                | -0.36                           | State government                                 |
| PCS1                      | -2.26                | -0.1                            | Utility Scale                                    |
| PCS2                      | 6.71                 | 0.13                            | Population-Service Connections Contrast          |
| PCD1                      | 5.36                 | -0.1                            | Economic Affluence                               |
| PCD2                      | 19.34                | -0.15                           | Wealthier, Urban Density & Rental                |
| PCD3                      | 1.66                 | 0.15                            | Homeownership Strength                           |
| PCD4                      | -4.07                | 0.08                            | Deep Poverty Gradient                            |
| PCD5                      | 9.36                 | -0.13                           | Demographic Homogeneity                          |
| PCD6                      | -2.48                | -0.02                           | Income Inequality                                |
| PCD7                      |                      |                                 | Median Income Variance                           |
| PCD8                      | 9.02                 | -0.09                           | Middle Income dominated                          |
| Health violations_10yr    | 0.85                 | X                               | 10 yr health violations                          |
| Av_water_charge           | X                    |                                 | Average Water charge                             |
| <b>Partial R2</b>         |                      |                                 |                                                  |

| Response →<br>Predictor ↓ | Average<br>Charge | 10 year<br>health<br>violations |
|---------------------------|-------------------|---------------------------------|
| State                     | 19.9              | 7.3                             |
| source                    | 0.4               | 2.7                             |
| owner                     | 1.3               | 0                               |
| Scale                     | 0                 | 0.1                             |
| Demographics              | 1.5               | 0.4                             |
| <b>Total</b>              | <b>23.1</b>       | <b>10.5</b>                     |

## D.4 Public:

### D.4.1 Executive Summary

When you combine: Municipal finance PCs, Operating expense pressure, Water source effects, Socioeconomic PCs, and State fixed effects, to look at the collective impact on financial and compliance performance, the high-level insights from the GLM analysis of public utilities are:

1. *Water source complexity drives compliance risk.*
2. *Financial stress (higher OpExp/Revenue) increases violations.*
3. *Municipal fiscal structure strongly influences water utility charges.*
4. *State regulatory and institutional context still exerts an independent effect.*
5. *Some states appear to be structurally underpricing water relative to compliance burden.*
6. *Some states price higher and show lower long-run violation deviations. However, higher state-level pricing deviations are not systematically associated with lower violations.*
7. *States with utilities with lower water charge heterogeneity tend to charge more on average; states with more variable pricing may have greater health violations.*

#### *Cross-Cutting Synthesis of State Effects:*

Note that these are the interpretations from the regression coefficients and NOT from the raw data – those were presented in the exploratory analysis section. The difference is that each statement here refers to how a predictor affects a response variable, while accounting for, i.e., controlling for all other predictors considered.

#### **Health and Total Violations Are Structurally Linked**

Very strong and consistent correlations:

- 5-year health  $\leftrightarrow$  10-year health:  $p = 0.82$
- 5-year total  $\leftrightarrow$  10-year total:  $p = 0.90$
- 5-year health  $\leftrightarrow$  5-year total:  $p = 0.64$
- 10-year health  $\leftrightarrow$  10-year total:  $p = 0.49$

#### **Interpretation**

There are persistent structural compliance environments at the state level: States that are high in short-run health violations are almost always high in long-run health violations. Total violations behave similarly.

#### *Oklahoma is a consistently high-violation state*

Across 5-year and 10-year health and total violations, Oklahoma stands out.

This indicates a *persistent structural compliance challenge* not explained by:

- Water source mix, Financial PCs, Socioeconomic PCs.

This is a true residual state effect.

#### *Arizona shows high long-run total violations and high charge variability*

Arizona stands out in:

- 10-year total violations (strongly positive), Std deviation of charges (positive).

This suggests:

- Structural regulatory stress, possibly high environmental treatment complexity, and a more variable rate design.

#### **Southeast states show lower charges but mixed compliance**

Southeast states (TN, AL, MS, GA, etc.) show:

- Lower average charges, some negative financial coefficients, Mixed violation signals.

This raises a policy-relevant pattern: Lower rates may coexist with structural compliance pressure confirming that operating expense pressure correlates with violations.

### **Northeast & Mid-Atlantic show higher charges but better long-run total compliance**

Some Mid-Atlantic states show:

- Higher charges, Lower long-run total violations.

This is consistent with:

Higher pricing supporting stronger compliance infrastructure.

### **Fiscal stress links to financial and compliance outcomes**

Georgia's higher odds of OpExp > Revenue, combined with some compliance deviations, suggest:

State fiscal policy environment matters even after adjusting for municipal finance PCs.

### *Detailed State Effect Results by Response Variable*

This section further elaborates on state-level differences in the responses of the key variables while accounting for the effects of the predictors.

#### **1. Average Water Charge (State Effects-Public Utilities)**

##### **Visual pattern**

- Southeast (TN, AL, MS, GA region) strongly lower than the reference state.
- Great Plains (KS, NE region) strongly negative.
- Washington and parts of the Mid-Atlantic (NJ/MD) are positive.
- Large regional gradient.

##### **Interpretation**

After adjusting for water source and finance structure:

- Southeast utilities tend to charge **less** than the reference state, Alaska.
- Some Northeastern states charge **more**.

This likely reflects:

- Regional rate regulation norms, Cost of service differences, Political tolerance for higher rates, Capital investment patterns.

Importantly:

***Lower average charges in the Southeast coincide with higher violation coefficients in some of those states — suggesting potential underpricing risk.***

#### **2. Standard Deviation of Water Charge (State Effects-Public Utilities)**

##### **Visual pattern**

- Arizona, Texas, Florida show high positive variability.
- Northeast (NJ/NY area) strongly negative.
- Midwest mostly neutral.

##### **Interpretation**

In some states, within-utility charge dispersion is much higher.

This could reflect: Tiered pricing structures, Urban-rural service variation, Rate class differentiation, Municipal cross-subsidization practices.

#### **3. 5-Year Total Violations (Public Utilities)**

##### **Visual pattern**

- Strong positive deviations: **Oklahoma, Arizona, Pennsylvania, West Virginia**
- Many other states near zero or mildly negative.

#### **Interpretation**

Even after adjusting for water source and financial structure, these states show higher total violation counts than the reference state.

This suggests: Regulatory enforcement intensity differences, Infrastructure age differences not captured predictors, Regional environmental stressors (e.g., groundwater contamination patterns), Institutional capacity variation.

### **4. 10-Year Total Violations**

#### **Visual pattern**

- Many states are strongly negative (blue), particularly across the Midwest, Southeast, and Northeast.
- Arizona stands out as strongly positive.
- Some Mid-Atlantic states show mild positive deviations.

#### **Interpretation**

This is the most striking map.

After controlling for covariates: Most states have lower long-run total violations than the reference. A few states (AZ in particular) have substantially higher long-run totals.

This suggests:

- The reference state, Alaska, may be relatively high-performing. Some states may have particularly strong regulatory frameworks. Arizona appears structurally high relative to national controls.

### **5. 5-Year Health Violations**

#### **Visual pattern**

- Oklahoma stands out strongly positive.
- Arizona and Louisiana also positive.
- Most other states near neutral.

#### **Interpretation**

Short-run health violations are highly concentrated geographically.

This implies:

- Specific regional compliance challenges, Possibly state-level enforcement/reporting differences, Or localized treatment complexity (e.g., arsenic, uranium, nitrates in Southwest).

The tighter clustering (fewer high states than total violations) suggests: Health violations may reflect more severe compliance stress than administrative violations.

### **6. 10-Year Health Violations**

#### **Visual pattern**

- Oklahoma remains strongly positive.
- Louisiana positive.
- Pattern is consistent with 5-year map but more persistent.

#### **Interpretation**

States elevated in 5-year health remain elevated in 10-year health.

This indicates **structural, persistent state-level risk**, not temporary fluctuations.

The persistence suggests:

- Long-term infrastructure deficits. Stable regulatory differences. Environmental baseline differences across states.

## **7. Operating Expense > Revenue (lopexrev, logit)**

### **Visual pattern**

- Georgia stands out as positive (higher odds of OpExp > Revenue).
- Many states are negative (lower odds of shortfall).
- Midwest and Mountain West generally slightly negative.

### **Interpretation**

State-level fiscal environments influence whether utilities run operating deficits.

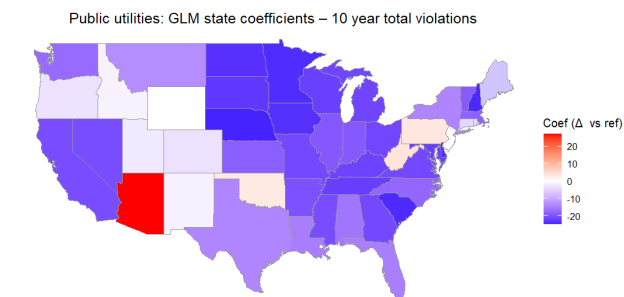
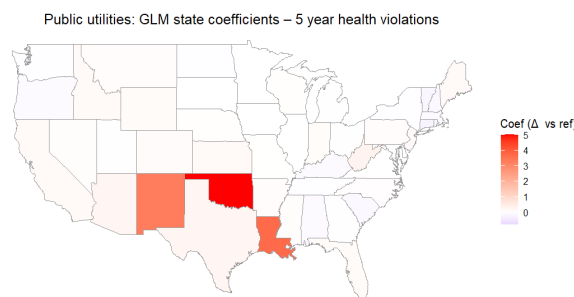
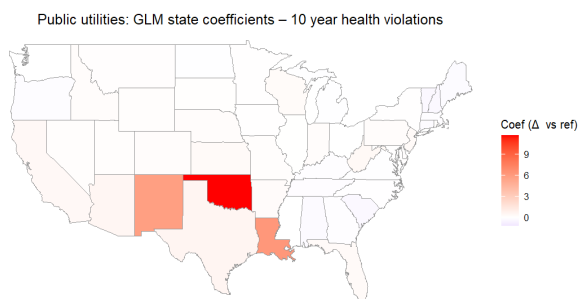
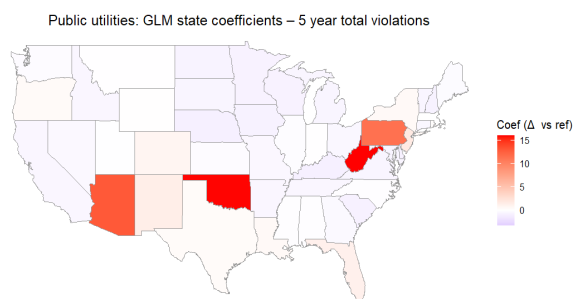
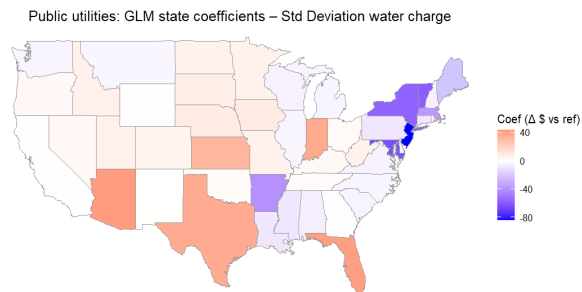
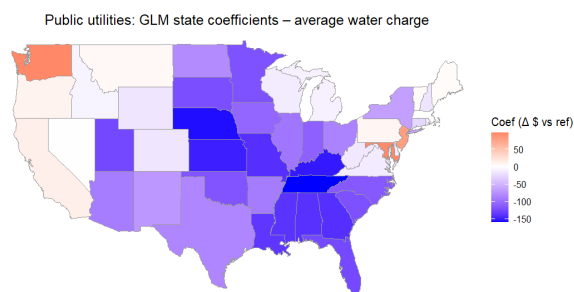
Georgia's positive coefficient suggests: Structural municipal fiscal stress. Rate regulation differences, Political constraints on pricing.

This is important because:

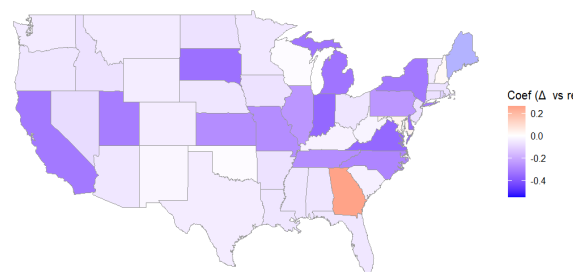
Operating expense pressure was also positively associated with violations in the earlier regressions.

### **This creates a possible chain:**

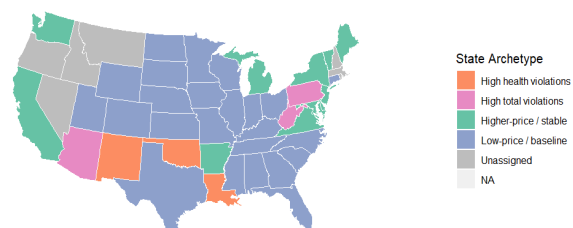
Municipal/State fiscal stress → utility shortfalls → elevated compliance risk.



Public utilities: GLM state coefficients – Operating Expense>Revenue Flag



State Archetype Clusters — Public Utilities  
Cluster membership from GLM state fixed-effect coefficients



Cluster 0 = Low-price baseline; 1 = High health violations; 2 = High total violations; 3 = Higher-price / stable

The states can be **clustered** into 4 groups based on these results (bottom right graphic above):

#### **Cluster 0 – Low-Price, Structurally Neutral Compliance (Baseline Majority)**

Strongly negative average charge deviations, Moderate compliance performance, No extreme fiscal stress, Large Midwest / Southeast grouping

#### **Cluster 1 – Persistent High Health Violations**

- Very high 5- and 10-year health violations, Elevated total violations, Pricing not proportionally higher, Structural compliance burden, Includes: OK, LA, NM

#### **Cluster 2 – High Total Violation Burden**

- Elevated long-run total violations, Moderate-to-high health violations, Mixed pricing structure, Includes: AZ, WV, PA

#### **Cluster 3 – Higher-Price / Financially Stable States**

- Higher or less-negative average charges, Lower long-run total violations, Lower rate variability, Generally stronger financial stability, Includes: CA, WA, NJ, NY, MD, VA, ME, MI, DE, VT, AR

#### D.4.2 GLM Results for all Public Regressions

Retaining only predictors that are significant at the 5% level.

| Response →<br>Predictor ↓ | Average<br>Charge | sdev<br>charge | 5 year<br>health | 10 year<br>health | 5 year<br>total | 10 year<br>total | lopexrev | Variable Definition                            |
|---------------------------|-------------------|----------------|------------------|-------------------|-----------------|------------------|----------|------------------------------------------------|
| source_GUP                |                   | 30             |                  |                   |                 |                  |          | Gr.water with surf. water influence, purchased |
| source_GW                 |                   |                |                  |                   | -5.7            | -5.9             |          | Ground Water                                   |
| source_GWP                | -25               | 15             | -0.6             |                   | -7.1            | -9.1             |          | Ground Water purchased                         |
| source_SW                 | -20               | 16             | 0.8              | 2.2               | -4.3            | -4.7             |          | Surface Water                                  |
| source_SWP                |                   | 17             |                  |                   | -6.9            | -9.0             |          | Surface Water Purchased                        |
| av_water_charge           |                   | 0.3            |                  |                   | 0.0             |                  |          | Av. Water Charge                               |
| sdev_water_charge         | 1.3               |                |                  |                   |                 |                  |          | StDev of water charge                          |
| health_viol_5yr           |                   |                |                  |                   |                 |                  |          | 5 year health violations                       |
| total_viol_5yr            | 0.3               | -0.2           |                  |                   |                 |                  |          | 5 year totlviolations                          |
| health_viol_10yr          |                   |                |                  |                   |                 |                  | 0.00     | 10 year health violations                      |
| total_viol_10yr           |                   |                |                  |                   |                 |                  |          | 10 year totatl violations                      |
| waterutilexp_rev          |                   |                | 0.2              | 0.6               |                 |                  |          | Op. Expense/Revenue Ratio                      |
| totalexpr_rev             |                   |                |                  | -0.5              |                 |                  | 0.14     | Total Expense/Revenue Ratio                    |
| lopexrev                  |                   |                |                  |                   |                 |                  |          | Is Utility Op Expense > Revenue?               |
| ltexprev                  |                   |                |                  |                   |                 |                  | 0.15     | Is Total Expense > Revenue                     |
| PSC1                      |                   | 3.9            | -0.1             | -0.1              |                 |                  |          | Population Scale                               |
| PSC2                      |                   |                |                  |                   |                 |                  |          | Population-Service Connections Contrast        |
| PCD1                      | -2.0              | 4.0            | -0.1             | -0.3              | -0.3            | -0.6             |          | Economic Stability                             |
| PCD2                      | -10               | 8.2            |                  | -0.2              | -0.5            | -0.7             | -0.02    | Urban Density & Rental Wealth                  |
| PCD3                      | 4.1               |                | -0.1             | -0.2              |                 |                  |          | The Vulnerability Gap                          |
| PCD4                      | 14                |                | 0.2              | 0.4               |                 |                  | 0.03     | Low-Income Homeownership                       |
| PCD5                      | -11               | 3.6            | -0.2             | -0.4              |                 |                  | -0.02    | Suburban Renters                               |
| PCD6                      |                   |                |                  |                   |                 |                  |          | Poverty Depth Gradient                         |
| PCD7                      | 10                |                |                  |                   |                 |                  |          | Income Inequality                              |
| PCD8                      |                   |                |                  |                   | -1.5            |                  |          | Occupancy & Poverty Variance                   |
| PCF1                      |                   | -1.5           |                  |                   |                 | -0.2             |          | System Financial Scale                         |
| PCF2                      | 3                 |                |                  |                   |                 |                  |          | Municipal vs. Utility Focus                    |
| PCF3                      |                   |                |                  |                   |                 | -1.2             |          | Capital Expansion vs. Tax                      |
| PCF4                      | 10                | -5.7           | 0.2              |                   |                 |                  |          | Investment Priority                            |
| PCF5                      |                   |                |                  |                   |                 |                  | 0.25     | Water utility Revenue expense contrast         |
| PCF6                      | -19               |                | -0.4             | -1.3              |                 |                  | -0.18    | Debt Service Pressure                          |
| PCF7                      | -49               | 19             |                  | -1.6              |                 |                  | -0.11    | General Debt Profile                           |
| PCF8                      |                   |                |                  |                   |                 |                  | 2.62     | Municipality Financial Viability               |
| PCF9                      |                   |                |                  |                   |                 |                  |          | Water Utility Expense Disaggregation           |
| PCF10                     |                   |                |                  |                   |                 |                  |          | Water Utility Capital deployment               |
| PCFS1                     | -2.2              | 1.1            |                  |                   |                 |                  |          | Muni revenue-Utility Tax                       |
| PCFS2                     |                   | -1.3           |                  |                   |                 |                  |          | Revenue/Expense Volume                         |
| PCFS3                     | -1.8              |                |                  |                   |                 |                  |          | Utility Cash flow                              |
| PCFS4                     | -4.7              | 3.4            |                  |                   |                 |                  | -0.01    | Muni Expenses on Infrastructure                |
| PCFS5                     | -9.2              | 4.4            |                  |                   |                 |                  | -0.04    | Expense                                        |
| PCFS6                     | -16               | 11             | -0.2             |                   |                 |                  | 0.07     | Expense Breakdown                              |
| PCFS7                     | -13               |                |                  |                   |                 |                  | -0.06    | Utility vs Municipality                        |
| PCFS8                     | -52               | 48             |                  |                   |                 |                  | 0.26     | Utility Tax dominates muni revenue             |
| PCFS9                     |                   |                |                  |                   |                 |                  |          | Debt vs Capital outlay                         |
| PCFS10                    | 11190             |                |                  |                   |                 |                  |          | Capital Project Nuance                         |

The fractional variance explained by each group of variables for each regression is reported in the table below. Note the positive correlation between the average and the standard deviation of the water charge. We see that state effects dominate all other effects.

| Response →<br>Predictor↓ | Average<br>Charge | sdev<br>charge | 5 year<br>health<br>violations | 10 year<br>health<br>violations | 5 year<br>total<br>violations | 10 year<br>total<br>violations | lopexrev    |
|--------------------------|-------------------|----------------|--------------------------------|---------------------------------|-------------------------------|--------------------------------|-------------|
| State                    | 19.6              | 16             | 10.6                           | 10.9                            | 8.4                           | 12.7                           | 6.1         |
| Water Source             | 0.1               | 0.6            | 1.8                            | 2.1                             | 0.5                           | 0.4                            |             |
| Av. Water Charge         |                   | 43.2           |                                |                                 | 0.1                           | 0.1                            |             |
| SD Water Charge          | 43.2              |                |                                |                                 |                               |                                |             |
| total_viol_5yr           | 0.1               | 0.1            |                                |                                 |                               |                                |             |
| health_viol_10yr         |                   |                |                                |                                 |                               |                                | 0.1         |
| total_viol_10yr          |                   |                |                                |                                 |                               |                                |             |
| waterutilexp_rev         |                   |                | 0.1                            | 0.2                             |                               |                                |             |
| totalexp_rev             |                   |                |                                |                                 |                               |                                | 0.7         |
| lopexrev                 |                   |                |                                |                                 |                               |                                |             |
| ltexprev                 |                   |                |                                |                                 |                               |                                | 1.5         |
| Scale                    |                   | 0.1            | 0.1                            |                                 |                               |                                |             |
| Demographics             | 1.8               | 2.9            | 0.9                            | 0.9                             | 0.3                           | 0.3                            | 0.3         |
| Finance                  | <u>1.3</u>        | <u>0.3</u>     | <u>0.1</u>                     |                                 |                               | <u>0.1</u>                     | <u>1.7</u>  |
| <b>Total</b>             | <b>66.1</b>       | <b>63.2</b>    | <b>13.6</b>                    | <b>14.1</b>                     | <b>9.3</b>                    | <b>13.6</b>                    | <b>10.4</b> |

#### D.4.3 Summary Analysis of Public GLM beyond state effects

In this sample of public utilities, municipal fiscal structure and finance indicators are the strongest and most consistent correlates of higher, more variable water charges and elevated odds of operating expense shortfalls. Surface-water systems and utilities under operating expense pressure are the clearest predictors of health and total violations. Socioeconomic context (low Economic Affluence, concentrated low-income homeownership, and inequality) compounds both financial strain and regulatory risk; together, these patterns point to both operational (treatment, O&M) and fiscal (municipal finance, rate structure) levers for intervention. Policy and assistance that target surface water systems and utilities embedded in fragile municipal fiscal environments is likely to deliver the biggest dividends for compliance and affordability.

##### 1. *Municipal fiscal environment & utility financing are central*

For average water charges, their heterogeneity, and utility financial distress (lopexrev), the most consistent, largest associations are with indicators that describe how a municipality raises revenue, its debt profile, and how expenses/revenue are structured between the municipality and the utility. In short, the municipal finance context matters — where it is stressed, it strongly accompanies higher charges and higher odds of operating expense shortfalls.

##### 2. *Water source is a dominant compliance driver*

Surface water systems (including purchased surface water) show consistently higher violation counts (both 5- and 10-year models) and are associated with greater variability in charges. This aligns with practical expectations: surface water often requires more complex treatment, so treatment/regulatory risk rise.

##### 3. *Operating expense pressure (expense/revenue) links finance and compliance*

The waterutilization\_rev ratio is repeatedly significant and positive in violation models. That is, utilities with tighter operating finances (higher operating expenses relative to revenue) have more violations. This suggests a financial → compliance link (e.g., insufficient O&M or deferred maintenance due to revenue shortfalls increases the risk of violations). Note this is associative, not causal.

##### 4. *Socioeconomic context shapes both finances and compliance*

Multiple demographic PCs (Economic Affluence, low-income homeownership, urban density/rental wealth, income inequality) are significant across models. Patterns include:

- Lower Economic Affluence (PCD1) is associated with higher violation counts.

- Low-income homeownership (PCD4) is associated with higher violations and in some models higher charges. These findings imply equity and affordability issues: communities with weaker socioeconomic profiles face both greater compliance risk and sometimes higher charges.

5. *Price outcomes and compliance outcomes are linked but distinct*

Charges and violations share predictors (municipal finance, socioeconomic PCs, water source), but the strongest signal of ownership/municipal finance is in pricing and price variability, while violations are most strongly explained by water source and operating expense pressure. In practice, some utilities pass compliance costs into rates, but compliance risk itself reflects operational and environmental constraints.

#### *D.4.4 Additional Insights from individual models:*

##### **1) Average Annual Water Charge**

###### **Managerial summary**

- Charges vary strongly with local fiscal structure and community socioeconomic status. The largest, most consistent associations are with municipal/utility finance indicators: **“Utility tax dominates muni revenue,” “General debt profile,” “Expense breakdown,”** and related municipal finance PCs.
- Socioeconomic components (e.g., **low-income homeownership, income inequality, urban density/rental wealth**) also matter and shift average charges meaningfully.
- Which municipality the utility sits in — its fiscal structure, debt profile, and how revenue is raised — is a dominant correlate of how much customers pay.

###### **Technical summary**

- Several finance-related PCs (PCFS8, PCF7, PCFS6, PCFS7, PCFS5, PCF4) are among the largest-magnitude significant coefficients. Some socioeconomic PCs (PCD2, PCD4, PCD5, PCD7) are also significant.

##### **2) Standard deviation (within-utility) of Charges**

###### **Managerial summary**

- Price variability within utilities (how uneven bills are across customers or accounts) is tightly linked to **water source** (surface & purchased surface/groundwater appear), and similar municipal finance indicators that drive average charge.
- Larger variability tends to appear in utilities embedded in municipalities with dominant utility taxes and particular expense structures.

###### **Technical summary**

- Source indicators (surface water, purchased surface/groundwater) and finance PCs (PCFS8, PCF7, PCFS6) are strongly positive predictors of s.d. of charges. Population scale and some socioeconomic PCs also retained significance.

##### **3) 5-year Health Violations (Poisson)**

###### **Managerial summary**

- **Surface-water systems** and utilities showing **operating expense pressure (Op Expense / Revenue)** are more likely to have higher health violations over 5 years.
- Community socioeconomic features (low-income homeownership, suburban renter profiles, vulnerability measures) modulate risk. Some socioeconomic features are associated with *higher* violation risk (e.g., low-income homeownership), while others (some stability indicators) are associated with *lower* risk.

###### **Technical summary**

- source\_SW (surface water) has a positive, significant coefficient. waterutilexp\_rev (Op expense/revenue ratio) is positive and significant. Several PCD components (PCD4 etc.) are significant, with mixed signs depending on the PC interpretation.

#### 4) 10-year Health Violations (Poisson)

##### Managerial summary

- The 10-year violation story reinforces the 5-year finding: **surface water** and **financial strain** (operating expense pressure) predict higher long-run violation counts.
- Socioeconomic structure still matters; areas with weaker Economic Affluence or higher low-income homeownership show higher long-run violation exposure.
- Operating expenses exceeding revenue is an indicator of higher violations.

##### Technical summary

- source\_SW has an even larger positive coefficient for 10-year violations than for 5-year. waterutilexp\_rev has a larger positive coefficient as well. PCD4 (low-income homeownership) positive; PCD1 (Economic Affluence) negative — consistent with the interpretation that weaker local economic conditions associate with more violations.

#### 5) 5-year total violations & 10-year total violations (Poisson)

##### Managerial summary

- Same pattern as the health-only violations: **water source and finance/expense pressure** are leading predictors of total compliance burden.
- Total violations follow the same socioeconomic gradients.

##### Technical summary

- Multiple finance and socioeconomic PCs are significant; total violation models often retain more predictors (larger # significant) because they combine health + other violations.

#### 6) lopexrev — Operating expense greater than revenue (binary; logit)

##### Managerial summary

- The strongest predictors of whether operating expense exceeds revenue are **municipal financial viability/municipality fiscal characteristics** and **utility tax dominance**; those municipal finance signals raise the odds that utility operating expense exceeds operating revenue.
- Debt service pressure and some debt profile components also relate to the likelihood of operating shortfall.

##### Technical summary

- PCF8 (Municipality Financial Viability) shows a strong positive coefficient (largest). PCFS8 (Utility tax dominance) and PCF5 (water utility rev/expense contrast) are also positive. PCF6 (debt service pressure) negative in sign (suggesting interpretation nuance). Because this is a logit coefficient, exponentiating gives odds ratios for interpretation.

## D.5 Private:

### *D.5.1 Executive Summary*

An interesting result from the analysis of the Private Utility data was that in most cases the differences across states were not statistically significant. The main exceptions were the states of KS, KY, NC, and NY, where average annual water charges are about \$400 lower than the reference, and TX for 5 and 10 year total violations that are higher than the reference. The finding that states play a less important role in the private utility regressions is an interesting outcome. It could well be that the smaller sample size for many states limits what is seen as statistically significant.

Private utility GLM results reinforce that regulatory compliance and pricing outcomes are primarily shaped by water source, financial structure, and state regulatory regimes.

### *D.5.2 Summary Analysis of Private GLM beyond state effects*

The cross regression observations are:

1. Operational margin (PCF3) is the strongest amplifier of long-run violation burden. Its regression coefficients are 3.6, 8.1, 100, and 208 for the 5 and 10 year health and total violations, respectively. Revenue-constrained private utilities face compliance challenges.
2. Groundwater systems mute violations; surface/mixed systems increase variability.
3. OpexRevratio increases pricing and compliance risk.
4. Affluent communities sustain higher prices and lower violations. Large Stable Households correspond to lower water charges but higher violations. Utilities that serve working-poor households have lower rates of charges and violations. Wealth disparity corresponds to a higher rate of heterogeneity.
5. Socioeconomic heterogeneity increases pricing dispersion.

### D.5.3 GLM Private Regression Coefficients

| Response →<br>Predictor↓ | Average Charge | sdev charge | 5 year health violations | 10 year health violations | 5 year total violations | 10 year total violations | lopexrev | Variable Definition                            |
|--------------------------|----------------|-------------|--------------------------|---------------------------|-------------------------|--------------------------|----------|------------------------------------------------|
| source_GUP               |                |             |                          |                           |                         |                          |          | Gr.water with surf. water influence, purchased |
| source_GW                | -137           |             |                          |                           |                         | -108                     |          | Ground Water                                   |
| source_GWP               |                |             |                          |                           |                         | -139                     |          | Ground Water purchased                         |
| source_SW                |                | 45          |                          |                           |                         | -104                     |          | Surface Water                                  |
| source_SWGW              |                |             |                          |                           |                         |                          |          | Surface water-groundwater                      |
| source_SWP               |                |             |                          |                           |                         |                          |          | Surface Water Purchased                        |
| av_water_charge          |                | 0.17        |                          |                           |                         |                          |          | Av. Water Charge                               |
| sdev_watercharge         | 1.5            |             |                          |                           |                         |                          |          | StDev of water charge                          |
| health_viol_5yr          |                |             |                          |                           |                         |                          |          | 5 year health violations                       |
| total_viol_5yr           |                |             |                          |                           |                         |                          |          | 5 year totlviolations                          |
| health_viol_10yr         |                |             |                          |                           |                         |                          |          | 10 year health violations                      |
| total_viol_10yr          |                |             |                          |                           |                         |                          |          | 10 year totatl violations                      |
| Opexprevratio            | 31             |             |                          |                           |                         |                          |          | Op. Expense/Revenue Ratio                      |
| totexprevratio           |                |             |                          |                           |                         |                          |          | Total Expense/Revenue Ratio                    |
| lopexrev                 |                |             |                          |                           |                         |                          |          | Is Utility Op Expense > Revenue?               |
| ltotopex                 | -42            |             |                          |                           |                         |                          |          | Is Total Expense > Revenue                     |
| PCS1                     |                | 4           |                          |                           |                         |                          |          | Population Scale                               |
| PCS2                     |                |             |                          |                           |                         |                          |          | Population-Service Connections Contrast        |
| PCD1                     | 30             | -12         |                          |                           |                         |                          |          | Economic Affluence                             |
| PCD2                     |                |             |                          |                           |                         |                          |          | Urban Diversity                                |
| PCD3                     |                | 11          |                          |                           |                         |                          | -0.03    | Knowledge Economy                              |
| PCD4                     |                | 18          |                          |                           |                         |                          |          | Educational Attainment                         |
| PCD5                     |                | -8          |                          |                           |                         |                          |          | Young Starter Families                         |
| PCD6                     | -35            |             |                          |                           | 9                       |                          |          | Large Stable Households                        |
| PCD7                     |                |             |                          |                           | 8                       |                          | 0.06     | Active Labor Force                             |
| PCD8                     |                | 8           |                          |                           |                         |                          |          | Fragile Asset-Holders                          |
| PCD9                     |                |             |                          |                           |                         |                          |          | Deep Poverty & Literacy                        |
| PCD10                    | -23            |             | -0.8                     |                           |                         |                          |          | The Working Poor                               |
| PCD11                    |                |             |                          |                           |                         |                          |          | Institutional Professionals                    |
| PCD12                    | -37            |             |                          |                           |                         |                          |          | Educational Residuals                          |
| PCD13                    |                |             |                          |                           |                         |                          |          | Vocational Mid-Income                          |
| PCD14                    |                |             |                          |                           |                         |                          |          | Established Urban Renters                      |
| PCD15                    |                |             |                          |                           |                         |                          |          | Immigrant Homeowners                           |
| PCD16                    | 45             | -17         |                          |                           |                         |                          |          | Educated Financial Stress                      |
| PCD17                    |                |             |                          |                           |                         |                          |          | Older Non-Professional Class                   |
| PCD18                    |                |             |                          |                           |                         |                          |          | Senior Lower-Middle Class                      |
| PCD19                    |                | 24          |                          |                           |                         |                          |          | Local Wealth Disparity                         |
| PCD20                    |                | ####        |                          |                           |                         |                          |          | Poverty-Middle Class contrast                  |
| PCF1                     |                |             |                          |                           |                         |                          |          | Financial Scale                                |
| PCF2                     |                |             |                          |                           |                         |                          |          | Tax vs Operating Expense Contrast              |
| PCF3                     |                |             | 3.6                      | 8.1                       | 100                     | 208                      |          | Operational Margin                             |
| PCF4                     |                |             |                          |                           |                         |                          |          | Overhead Impact                                |
| PFS1                     |                |             |                          |                           |                         |                          |          | Financial Intensity                            |
| PFS2                     |                |             |                          |                           |                         |                          | 0.07     | Tax/Expense Profile                            |
| PFS3                     |                |             |                          |                           |                         |                          | -0.18    | Operational Efficiency                         |
| PFS4                     |                |             |                          |                           |                         |                          |          | Revenue Surplus                                |

The partial variance explained by each variable (group) is presented below. As with the public utilities, the average and standard deviation of water charge are correlated. Demographics appear as a much more important variable here, with economic affluence being a dominant variable.

#### Partial R2

| Response →<br>Predictor ↓ | Average<br>Charge | sdev<br>charge | 5 year<br>health<br>violations | 10 year<br>health<br>violations | 5 year<br>total<br>violations | 10 year<br>total<br>violations | lopexrev  |
|---------------------------|-------------------|----------------|--------------------------------|---------------------------------|-------------------------------|--------------------------------|-----------|
| State                     | 0                 | 0              | 0                              | 0                               | 20                            | 22                             | 14        |
| Source                    | 0                 | 4              | 0                              | 0                               | 8                             | 6                              | 3         |
| av_water_charge           | 0                 | 28             | 0                              | 0                               | 0                             | 0                              | 0         |
| sdev_watercharge          | 29                | 0              | 0                              | 0                               | 0                             | 0                              | 0         |
| health_viol_5yr           | 0                 | 0              | 0                              | 0                               | 0                             | 0                              | 0         |
| total_viol_5yr            | 0                 | 0              | 0                              | 0                               | 0                             | 0                              | 0         |
| health_viol_10yr          | 0                 | 1              | 0                              | 0                               | 0                             | 0                              | 0         |
| total_viol_10yr           | 0                 | 0              | 0                              | 0                               | 0                             | 0                              | 0         |
| Opexprevratio             | 2                 | 0              | 0                              | 0                               | 0                             | 0                              | 0         |
| totexprevratio            | 0                 | 0              | 0                              | 0                               | 0                             | 0                              | 0         |
| lopexrev                  | 0                 | 0              | 0                              | 0                               | 0                             | 0                              | 0         |
| ltotopex                  | 1                 | 0              | 0                              | 0                               | 1                             | 0                              | 0         |
| Population Scale          | 0                 | 0              | 6                              | 13                              | 0                             | 24                             | 0         |
| demographics              | 31                | 59             | 4                              | 0                               | 6                             | 1                              | 3         |
| <u>financial</u>          | <u>1</u>          | <u>2</u>       | <u>15</u>                      | <u>10</u>                       | <u>22</u>                     | <u>32</u>                      | <u>3</u>  |
| <b>Total</b>              | <b>64</b>         | <b>94</b>      | <b>25</b>                      | <b>24</b>                       | <b>57</b>                     | <b>84</b>                      | <b>24</b> |

#### D.5.4 Response-by-Response Summary (Private Utilities)

##### 1. Average Water Charge

**Positive Drivers (Increase Charges):** *Opexprevratio* (Operating Expense / Revenue Ratio) , *PCD1* - Economic Affluence, *PCD16* – Educated Financial Stress.

**Negative Drivers (Decrease Charges):** *source\_GW* (Groundwater), *ltotopex* (Total Expense > Revenue), *PCD6* – Large Stable Households, *PCD10* – Working Poor, *PCD12* – Educational Residuals.

*Interpretation:* Although higher operating expense pressure (*Opexprevratio*) is associated with higher water charges—consistent with cost pass-through pricing—utilities operating in total deficit (*ltotopex*) exhibit lower average charges. This apparent contradiction reflects two distinct financial regimes. *Opexprevratio* captures marginal operating cost pressure in systems that retain pricing flexibility, whereas *ltotopex* captures structural fiscal distress, often in economically constrained communities where rate increases are politically or socially limited. Thus, moderate cost pressure leads to upward price adjustment, while deeper fiscal imbalance reflects underpricing relative to cost and constrained financial capacity rather than aggressive rate-setting.

##### 2. Standard Deviation of Charges

**Positive Drivers (More Rate Dispersion):** *source\_SW* , *PCS1* – Population, *PCD4* – Educational Attainment, *PCD3* – Knowledge Economy , *PCD8* – Fragile Asset-Holders, *PCD19* – Local Wealth Disparity , *PCD20* – Poverty Middle Class Contrast

**Negative Drivers (Less Dispersion):** *PCD1* – Economic Affluence , *PCD5* – Young Starter , *PCD16* – Educated Financial Stress

*Interpretation:* Socioeconomic heterogeneity increases rate dispersion. Affluent and demographically stable communities exhibit more uniform water charges.

##### 3. 5-Year and 10 year Health Violations

**Positive Drivers:** *PCF3 – Operational Margin*

**Negative Drivers:** *PCD10 – Working Poor*

*Interpretation:* Operational margins amplify health violations, while areas with working poor seem to lower

#### **4. 5-Year and 10-year Total Violations**

**Positive Drivers:** *PCF3 – Operational Margin*

**Negative Drivers:** *source\_GW / GWP/SW for 10-year only*

*Interpretation:* Operational margin dominates total violation burden.

#### **5. lopexrev (Operating Expense > Revenue)**

**Positive Drivers:** *PCD7 – Active Labor Force, PFS2 – Tax/Expense Profile*

**Negative Drivers:** *PCD3 – Knowledge Economy, PFS3 – Operational Efficiency*

*Interpretation:* The lopexrev specification indicates that structural operating deficits are associated with communities characterized by an active labor force (PCD7) and specific tax/expense configurations (PFS2), while knowledge-economy characteristics (PCD3) and operational efficiency (PFS3) reduce the probability of operating expense exceeding revenue. These results suggest that utilities serving economically active but not necessarily affluent populations face higher operating intensity and fiscal strain, particularly when embedded within fiscal architectures that allocate costs unfavorably. In contrast, communities with higher human capital and professionally oriented economic structures appear to sustain stronger revenue bases, while operational efficiency directly mitigates deficit risk. The findings highlight that operating imbalance is jointly shaped by community economic composition, fiscal design, and managerial efficiency.

## D6. Overall Comparative Conclusions at a Glance:

### Pervasive Predictors Across Models

Below are the domains that consistently influence multiple responses.

#### 1. Environmental Complexity (Water Source)

##### Responses influenced:

- 5-year health violations ↑
- 10-year health violations ↑
- 5-year total violations ↑
- 10-year total violations ↑
- Often ↑ charge variability
- Sometimes ↑ average charges

##### Direction:

Surface water → **amplifies compliance burden**

Groundwater → often **mutes compliance burden**

**Interpretation:** Strongest cross-model consistent driver of violations.

#### 2. Financial Stress / Operational Margin

##### Responses influenced:

- Health violations ↑
- Total violations ↑
- Sometimes ↑ pricing
- ↑ probability expense > revenue

##### Direction:

Higher expense pressure → **amplifies violations**

Stronger operational margin (private) → appears positively associated with

**Interpretation:** Financial fragility increases compliance risk across ownership.

#### 3. State Institutional Context

##### Responses influenced:

- All violation measures
- Pricing
- Charge variability
- Financial distress

##### Direction:

State-dependent; clusters amplify or mute all outcomes.

**Interpretation:** Institutional regime acts as background structural multiplier.

#### 4. Municipal Fiscal Structure (Public Only)

##### Responses influenced:

- Average charge ↑ or ↓ depending on tax structure

- Charge variability
- Sometimes violation burden indirectly

**Direction:**

Greater municipal fiscal dependence → amplifies pricing dispersion

**Interpretation:** Public utilities are embedded in municipal fiscal architecture.

## 5. Socioeconomic Structure

**Responses influenced:**

- Health violations
- Total violations
- Pricing (sometimes)
- Financial stress probability

**Direction patterns:**

- Higher poverty / working poor → amplifies violations
- Economic affluence → often mutes violations
- Inequality → mixed, sometimes amplifies pricing

**Interpretation:** Community structural vulnerability affects compliance resilience.

## D7: Technical Influence Matrix

Below is a domain-level influence matrix summarizing pervasive effects.

Legend:

+= Amplifies

– = Mutes

0 = weak / inconsistent

++ = strong pervasive effect

| Structural Domain                   | Avg Charge | SDev | Health5 | Health10 | Total5 | Total10 | Lopexrev |
|-------------------------------------|------------|------|---------|----------|--------|---------|----------|
| Environmental (Surface Water)       | +          | +    | ++      | ++       | ++     | ++      | 0        |
| Financial Stress / Margin           | +          | 0    | +       | +        | +      | +       | ++       |
| State Institutional Context         | +          | +    | ++      | ++       | ++     | ++      | +        |
| Municipal Fiscal Structure (Public) | ++         | +    | 0/+     | 0/+      | 0      | 0       | ++       |
| Socioeconomic Vulnerability         | +          | 0    | +       | +        | +      | +       | +        |

## D8. Amplifiers vs. Muters

**Strong Amplifiers of Violations**

- Surface water dependence
- Financial stress (expense pressure)
- Certain states
- High poverty / working poor PCs

**Muters of Violations**

- Economic affluence

- Groundwater systems
- Some higher-price regulatory regimes

#### **Strong Amplifiers of Pricing**

- Municipal fiscal structure (public)
- Regulatory regime (private)
- Financial stress

## **D8 Geographic and Demographic Structure of Utility Performance**

### **1. State-Level Geography and Persistent Regulatory Regimes**

Across the public, private, and combined (Emma) models, state fixed effects remain large and statistically significant predictors of both compliance and pricing outcomes after controlling for environmental complexity, financial structure, and socioeconomic covariates. This persistence indicates that water utility performance is geographically structured.

In both public and private specifications, certain states exhibit systematically elevated 5-year and 10-year health violations (e.g., OK, LA, NM, WV in public; OK, LA, NM, NC in private). These positive state coefficients (amplifying violation burden) persist even when controlling for water source, financial stress, and demographic composition. Conversely, other states exhibit consistently negative coefficients (muting violations).

This geographic clustering suggests that:

- Regulatory enforcement intensity varies across states,
- Institutional capacity differs regionally,
- Environmental baseline conditions (e.g., groundwater contamination in the Southwest, legacy industrial contamination in Appalachia) create persistent compliance burdens.

This finding aligns with institutional analyses emphasizing regulatory regime heterogeneity (Levin & Tadelis, 2010; Shleifer, 1998) and with water governance scholarship documenting state-level variation in enforcement and oversight (Teodoro, 2018).

Importantly, state-level effects also influence pricing. In private models, average charge coefficients vary dramatically by state (e.g., strong negative deviations in KS, KY, NC; positive deviations in TX, VA), suggesting that rate approval regimes and political economy of pricing differ geographically.

**Implication:** Water utility performance is regionally embedded. Policy interventions should account for state regulatory culture and environmental baseline rather than treating utilities as isolated administrative units.

### **2. Population Scale and System Size Effects**

Population scale (e.g., PCS1 in the public models, and analogous size components in other specifications) shows systematic associations with violations and sometimes with charge variability.

In the public models:

- Larger population scale (PCS1) is associated with reductions in health violations in some specifications (negative coefficient), suggesting scale economies in compliance capacity.
- However, charge variability may increase with size in certain states, indicating more complex rate structures in larger systems.

In the private models:

- Operational margin and financial scale (e.g., PCF1/PCF3 domains) influence total violations and pricing.
- Larger systems appear more financially structured but not necessarily insulated from compliance burden.

These patterns are consistent with literature on scale economies in water utilities (Estache & Rossi, 2002), which finds that larger systems often benefit from technical and managerial scale efficiencies but may also face more complex regulatory oversight.

**Implication:** System size affects both compliance capacity and financial structure. Small, fragmented systems may be more vulnerable to financial stress and compliance risk, especially in economically heterogeneous regions.

### 3. Socioeconomic Heterogeneity and Community Vulnerability

A striking and consistent pattern across all three analyses is the influence of socioeconomic principal components.

While PCs differ by dataset, grouping by definition reveals the following structural relationships:

#### A. Economic Affluence (PCD1-type components)

- Higher economic affluence is generally associated with lower health violations (negative coefficients in public and private models).
- Affluent communities are better positioned to absorb compliance costs and maintain infrastructure.

This aligns with environmental justice literature showing that wealthier communities exhibit better compliance outcomes (Switzer & Teodoro, 2018; Teodoro, 2018).

#### B. Poverty and Working-Class Structure (e.g., PCD10 – “The Working Poor”; PCD4 – Educational Attainment; PCD19 – Local Wealth Disparity)

In private models:

- “The Working Poor” (PCD10) shows a negative coefficient on pricing (–23) and a negative coefficient on health violations (–0.8 for 5-year health), indicating muted pricing and potentially muted health burden in some contexts.
- However, “Local Wealth Disparity” (PCD19) shows positive effects on charge dispersion (e.g., +24 in sdev), indicating greater intra-system inequality in pricing structures.

In public models:

- Poverty-related components are associated with increased health violations and sometimes increased total violations.

The combined pattern suggests that economically vulnerable communities may experience:

- Lower average charges (muted pricing),
- But elevated or structurally persistent compliance burden,
- And greater heterogeneity in rate design.

This is consistent with the “affordability–compliance tension” documented in water governance literature (Teodoro, 2018).

**Implication:** Economically heterogeneous communities face structurally different risk profiles, where fiscal capacity and compliance burden interact in complex ways.

### 4. Rich vs. Poor Communities: Amplifiers and Muters

Across models:

| Structural Factor      | Violations             | Pricing                  | Interpretation                                             |
|------------------------|------------------------|--------------------------|------------------------------------------------------------|
| Economic Affluence     | – (mutes)              | + or neutral             | Wealth supports compliance capacity                        |
| Poverty / Working Poor | + (amplifies) or mixed | – (muted pricing)        | Vulnerability increases risk but constrains rate increases |
| Income Inequality      | + for variability      | + for pricing dispersion | Greater internal heterogeneity                             |

The key structural tension is:

Poorer communities often exhibit either muted pricing or elevated violation burden, reflecting limited fiscal capacity and infrastructure constraints.

This pattern reinforces environmental justice findings that disadvantaged communities face disproportionate regulatory risk (Switzer & Teodoro, 2018).

## **5. Regional Patterns Across the United States**

Combining state fixed effects and demographic components reveals broad regional structure:

### **Southwest (AZ, NM, TX)**

- Elevated long-run total violations (TX extreme in private model)
- Higher charge variability
- Environmental stress likely dominant (e.g., salinity, arsenic)

### **Appalachia (WV, KY)**

- Elevated violation burden in public models
- Economic vulnerability components relevant

### **Southeast (AL, MS, NC)**

- Lower average pricing in public models
- Mixed compliance signals
- Municipal fiscal structure important

### **Northeast / Mid-Atlantic (NJ, NY, MD)**

- Higher pricing in some models
- Lower long-run total violations
- Greater regulatory oversight and capital investment

These patterns indicate that:

Geography interacts with socioeconomic structure and regulatory regime to produce persistent spatial clustering of both pricing and compliance outcomes.

## **6. Implications for Economically Heterogeneous Communities**

1. Poorer communities may face elevated compliance risk but lack pricing capacity to fund remediation.
2. Wealthier regions appear better able to convert pricing flexibility into compliance resilience.
3. State-level regulatory environment may amplify or mitigate local structural vulnerability.
4. Regional targeting of capital grants should incorporate both environmental complexity and socioeconomic vulnerability metrics.

## D9 Technical appendix (GLM results)

### D9.1 Data and model specification

- Responses:
  - Average customer water charge (continuous, modeled with Gaussian GLM).
  - Standard deviation of customer charges within each utility (continuous, Gaussian).
  - Indicator that operating expenses exceed revenues (lopex, binomial GLM).
  - Health-based violation counts over 5 and 10 years (counts, modeled with Poisson GLM).
- Total violations over 5 and 10 years (counts, Poisson).
- Predictors (public and private, with separate PCA bases):
  - State fixed effects and primary source categories (e.g., groundwater, surface water, purchased water).
  - Demographic PCs (PCD): summarize socioeconomic advantage, poverty, education, age structure, homeownership, racial composition, and employment patterns, with separate component structures for public and private utilities.
  - Financial PCs (PCF): capture overall financial scale, contrasts between revenue and expenses, capital versus current spending, and debt and tax patterns.
  - Financial-structure-per-capita PCs (PCFS): express similar financial contrasts on a per-capita basis.
  - System size PCs (PSC): measure average scale of population served and service connections and the contrast between them.
  - Additional utility-level covariates: total expenditures to revenue ratios, historical violation counts, and related metrics as specified in each model.

### D9.2 Model performance overview

| Outcome (public)           | N    | $R^2$ / pseudo- $R^2$ | RMSE / scale | Key notes                                                                                                 |
|----------------------------|------|-----------------------|--------------|-----------------------------------------------------------------------------------------------------------|
| Average charge             | 8007 | ~0.62                 | ~114         | Strong explanatory power; state, sdev of charges, and multiple PCD/PCF/PCFS terms contribute.             |
| Sdev of charges            | 8007 | ~0.63                 | ~57          | Strong model; influenced by mean charge, source, and several demographic and financial components         |
| Lopex (opex > revenue)     | 8007 | ~0.16                 | ~0.45        | Modest fit; financial scale and structure are more important than demographics.                           |
| Health violations, 5-year  | 8007 | ~0.16                 | ~3           | Geography and source dominate; selected demographic and financial PCs have small but significant effects. |
| Health violations, 10-year | 8007 | ~0.17                 | ~6           | Similar to 5-year model with slightly higher explained variance.                                          |
| Total violations, 5-year   | 8007 | ~0.11                 | ~14          | Lower explanatory power; state and source explain most of what is captured.                               |
| Total violations, 10-year  | 8007 | ~0.15                 | ~23          | Geography remains dominant; some financial PCs contribute.                                                |

| Outcome (private)          | N   | $R^2$ /pseudo- $R^2$ | RMSE / scale | Key notes                                                                                                                                                     |
|----------------------------|-----|----------------------|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Average charge             | 277 | ~0.63                | ~135         | Much of the explained variance comes from state, primary source, and private PCD/PCF terms that encode socioeconomic advantage and revenue–expense contrasts. |
| Sdev of charges            | 277 | ~0.67                | ~44          | Influenced by private per-capita financial PCs and demographic segmentation.                                                                                  |
| Lopex                      | 277 | ~0.23                | ~0.34        | Driven by private-side financial-structure contrasts; partial $R^2$ is concentrated in a few PCs.                                                             |
| Health violations, 5-year  | 277 | ~0.26                | ~2.8         | Models have limited explanatory power; state/source and a small set of PCD/PCF components matter.                                                             |
| Health violations, 10-year | 277 | ~0.25                | ~7.7         | Similar pattern to 5-year model with small contributions from financial status.                                                                               |
| Total violations, 5-year   | 277 | ~0.46                | ~58          | State and a few PCs explain modest variance in violation counts.                                                                                              |
| Total violations, 10-year  | 277 | ~0.66                | ~92          | Geography and private financial structure remain the main drivers.                                                                                            |

#### D9.3 Key coefficient patterns (public)

- **Average charge (Gaussian)**

- Many states have significant negative deviations from the reference state, indicating lower average charges, while a few (for example, states with high infrastructure or scarcity costs) show positive deviations.
- Higher within-utility charge variability is positively associated with mean charge, suggesting that systems with more heterogeneous rate structures tend to have higher typical bills.
- Demographic PCs:
  - Components representing **social affluence and higher income** (for example, PCD4, PCD7 for public) are positively associated with charges, while PCs representing high renters or high poverty relate to lower charges.
- Financial PCs:
  - PCs indicating **large financial scale or high capital outlay** are associated with higher average charges, consistent with rate-setting to recover investment costs.

- **Sdev of charges (Gaussian)**

- The standard deviation increases with higher mean charges and in systems using certain primary sources, especially where surface water and complex financial structures coincide.
- Demographic PCs that reflect mixed or polarized communities (for example, high renters with low vehicles, or high poverty with homeowners) contribute to greater heterogeneity, implying more uneven rate burdens within those utilities.

- **Lopex (binomial)**

- Coefficients on total expenditures and lagged expenses are positive and significant, showing that high-spending utilities are more likely to have operating expenses exceed revenues.
- Financial PCs that combine high current expenses, high debt, or large capital programs tend to increase lopex risk, while PCs indicating strong revenue with low debt or lower opex are protective.
- **Health and total violations (Poisson)**
  - Several states have significantly higher expected violation counts relative to the reference, even after controlling for other factors, reflecting spatial clustering of regulatory risk.
  - Primary surface water sources are associated with higher health-violation rates compared to groundwater in the public models.
  - Specific demographic PCs related to poverty and insurance coverage have negative coefficients in some health-violation models (suggesting better performance in more affluent, insured communities) and positive coefficients where high poverty or unstable demographics are present.

#### *D9.5 Key coefficient patterns (private)*

- **Average charge and sdev of charges**
  - Private utilities show large state and source effects, but within that, PCs capturing **socioeconomic advantage, high middle incomes, and wealthy segments** are associated with higher average charges.
  - Charge heterogeneity is related to per-capita financial status and contrasts between revenue and expenses per capita, indicating that differences in how private utilities allocate costs drive dispersion in customer bills.
- **Lopex**
  - Private lopex models highlight PCs where expenses and taxes dominate revenues, confirming that specific financial structures, rather than demographic mix, are the primary predictors of operating deficits or near-deficits.
- **Violations**
  - Private violation models have fewer significant PCs; state and source type remain the strongest predictors, but certain high-poverty or high-unemployment PCs are associated with elevated total violations.

## Appendix E Random Forest Models

The random forest models explore the same data sets as the GLM models presented in Appendix D. The primary difference is that instead of having a linear model, the random forest models look for nonlinear effects. These may manifest in multiple ways, e.g., the frequency of violations increases dramatically if the median household income falls below some threshold, or certain combinations of conditions, e.g., violations occur only if the utility has high operating expense per person served, has surface water as a source, is publicly owned, and the location has low home ownership. These distinctions do not emerge easily with linear models, since they assume proportional responses to each predictor across all conditions. In this section, we illustrate that often the Random Forest models are able to explain or predict the response variables better than the GLM models and that we can extract more interpretable explanations by looking graphically at the relationship between the response variable and the individual predictor, as well as pairs of predictors, or how a combination of conditions can explain the difference in the response between public and private utilities.

For Random Forest models, three types of measures are typically reported. The first is performance in terms of  $R^2$  (variance explained), reported for both the data set used to fit the model (training) and the out-of-sample data used to test the model's predictions (CV). The second is feature importance, which indicates how much of the overall variability in the response variable is explained by each predictor. The third is the SHAP, which measures the contribution of each predictor to explaining each observation of the response, relative to the explanation of that observation by all the predictors. Plots are provided illustrating the SHAP values for all observations retained for model testing, and can also be provided for the average output. Next, one can also derive how the response variable depends on each predictor (marginal plots) or each pair of predictors (heat maps or contour plots) to see the nature of the relationship as predicted by the full model. We present selected examples of each dependence plot, along with the importance and SHAP plots for each Random Forest Model application, and then offer a synthesis of the implications for each analyzed data set. In all cases, we note that the variance explained is higher than for the linear GLM models.

### E1. Emma

We first consider the Emma data set, which has the largest number of utilities and covers the SDWIS and ACS data but lacks financial data.

E1.1 Average Water Charge

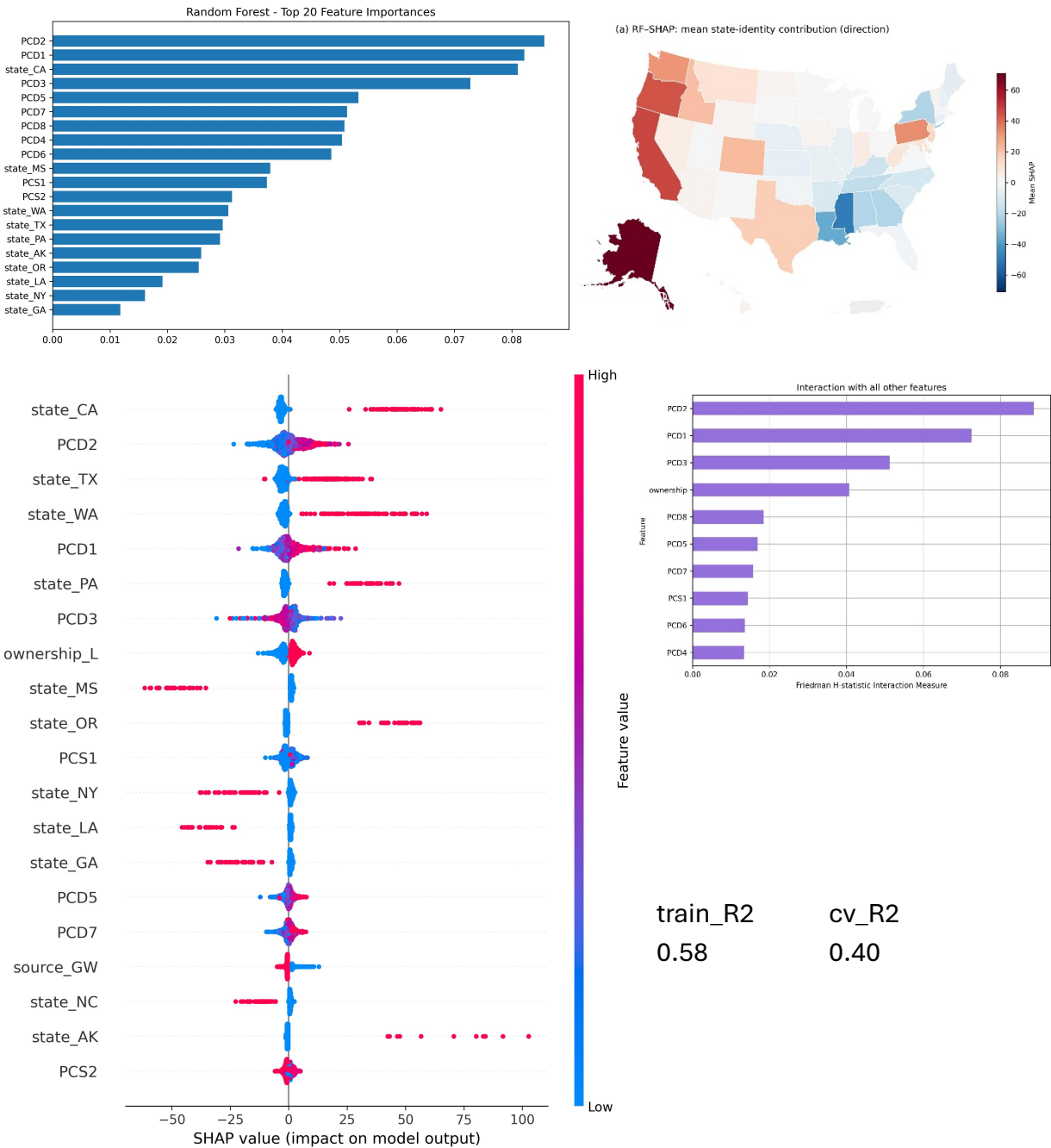


Figure E1. Random Forest Model for Predicting the Average Water Charge at the Utility Level for the EMMA data set

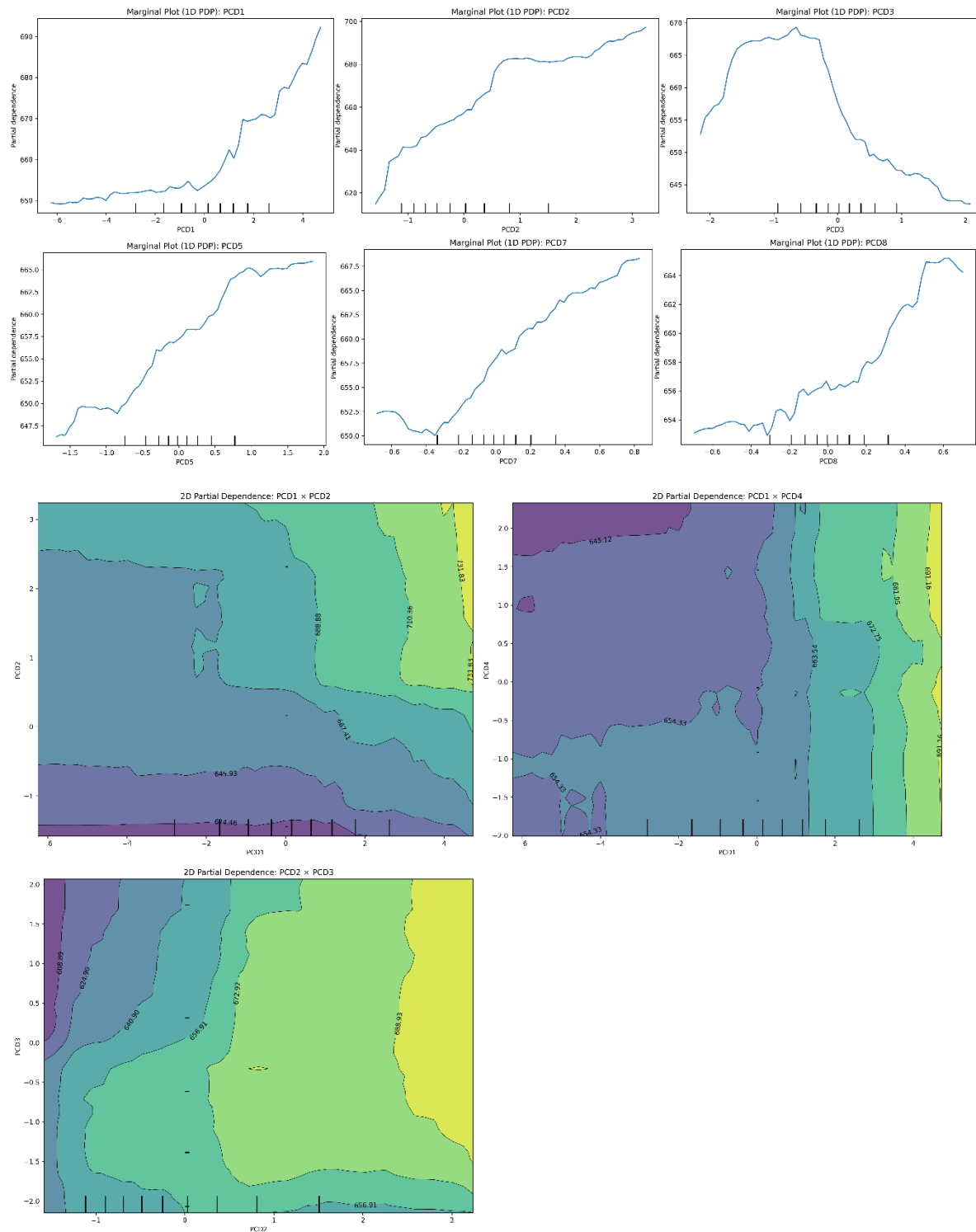


Figure E2. Partial dependence plots for selected variables from the Random Forest Model for Predicting the Average Water Charge at the Utility Level for the EMMA data set

The model incorporates demographic principal components (PCD1–PCD8), Scale Principal Components (PCS1–PCS2), state identity indicators, utility ownership type, and water source variables. We synthesize Figures 1 and 2, which contain feature importance rankings, SHAP values, partial dependence plots (1D and 2D), and interaction statistics, to characterize the principal drivers of water charges and the nature of their relationships—linear versus nonlinear, marginal versus interactive.

The analysis reveals that **socioeconomic and demographic structure is the dominant driver of variation in water pricing, with state-level geographic effects serving as a strong secondary layer**. Critically, most relationships are substantially nonlinear, with threshold effects, saturation plateaus, and cross-feature interactions all playing important roles.

*Individual Features:*

The top-ranked features by mean decrease in impurity (MDI) are, in descending order:

| Rank | Feature  | Interpretation                     |
|------|----------|------------------------------------|
| 1    | PCD2     | Wealthier, Urban Density & Rental  |
| 2    | PCD1     | Economic Affluence                 |
| 3    | state_CA | California indicator               |
| 4    | PCD3     | Homeownership Strength             |
| 5    | PCD5     | Demographic Homogeneity            |
| 6    | PCD7     | Median Income Variance             |
| 7    | PCD8     | Middle Income Dominated            |
| 8    | PCD4     | Deep Poverty Gradient              |
| 9    | PCD6     | Income Inequality                  |
| 10   | state_MS | Mississippi indicator              |
| 11   | PCS1     | Population and Service Connections |
| 12   | PCS2     | Population and Service Contrast    |

PCD2 is the single most important feature, followed closely by PCD1, together accounting for substantially more predictive variance than any other variables. The prominence of multiple demographic PC dimensions signals that **charge levels are not driven by a single socioeconomic axis but by a composite structure** spanning income gradients, urban character, homeownership, and poverty depth.

The SHAP beeswarm plot reinforces the importance ranking while adding directional information. Key observations:

- **state\_CA** shows the widest rightward SHAP spread: California utilities consistently charges substantially higher than the baseline. This is the strongest positive state effect observed.
- **PCD2** (Wealthier, Urban Density & Rental ) exhibits strong positive SHAP values at high feature values (pink/red dots extending to +75–+100), confirming that urbanized, wealthier service areas face higher charges. A strong, rapid ramp in the –1.5 to +0.8 region is followed by near-saturation above PCD2 = +1. This is a classic **sigmoidal/saturating response**: the marginal effect of additional urban wealth and diversity on charges diminishes once a high urbanization threshold is crossed. PCD2 is the most

powerful predictor. *The steep ramp indicates that shifting from low to moderate urbanization/wealth has an outsized effect on charges, more so than any further increase beyond that point.* The saturation implies that the most urban, wealthiest markets have already converged to a high-charge equilibrium where further demographic richness may not translate to additional price escalation.

- **PCD1** (Economic Affluence) shows concentrated positive SHAP at high values, with broad negative spread at low values—consistent with a monotonically increasing PDP relationship spanning roughly 650 to 692. *The curve is largely flat from PCD1 = −7 to approximately 0, suggesting that charges are relatively insensitive to variation in Economic Affluence within the below-average range.* Above zero, the relationship accelerates steeply, with the most rapid increase occurring between PCD1 = +1 and +4.5. This implies a **threshold effect**: *Economic Affluence must exceed a baseline before it meaningfully drives up charges, after which the relationship is strongly positive. Utilities serving economically affluent populations charge more, consistent with greater willingness-to-pay and potentially higher service quality expectations. The flat lower tail suggests that economically depressed areas are clustered at a low-charge floor, not differentiated from one another.*
- **state\_TX** and **state\_WA** both show right-skewed positive SHAP distributions, indicating they systematically elevate charges above the national average.
- **state\_MS** is a clear negative outlier: Mississippi utilities receive strongly negative SHAP values (dots extending to −50 and beyond), reflecting characteristically low charges.
- **state\_AK** (Alaska) shows a bimodal distribution with several extreme positive outliers, likely reflecting the heterogeneity of Alaskan utility systems—some with very high remote service costs.
- **PCD3** (Homeownership Strength) shows a mixed SHAP distribution, consistent with the non-monotonic PDP shape. The inverted-U shape is the most complex marginal relationship in the model. The peak at moderately negative PCD3 values (i.e., moderate homeownership reduction) is counterintuitive and suggests the relationship is genuinely interactive—the interpretation of PCD3 almost certainly depends on which other demographic axes are active. Communities with intermediate homeownership profiles (neither the most renter-heavy nor the most owner-dominated) tend to face the highest charges. This could reflect mixed urban/suburban contexts where infrastructure costs are high. The decline at high homeownership values may represent lower-density, lower-cost rural areas with owner-occupied single-family housing.
- **PCD5** (Demographic Homogeneity) has a monotonically increasing with a distinct plateau between approximately PCD5 = −1.2 and −0.8. The Range of effect is ~646 to ~666. The early plateau followed by a resumption of increase suggests two regimes: *demographically heterogeneous areas at low values cluster at moderate charges, while demographically homogeneous areas at high values pay notably more.* The step-like feature around PCD5 = −0.5 suggests a possible threshold beyond which homogeneity begins to drive charges upward more steeply. *Demographically homogeneous communities tend to face higher average charges. This may reflect that such communities are often suburban or exurban, with higher per-connection infrastructure costs and fewer economies of scale.*
- **PCD7** (Median Income Variance) Broadly monotonically increasing with a noticeable flattening/reversal near PCD7 ≈ −0.4 before resuming increase. Range of effect: ~650 to ~668 (approximately +18 units). A local minimum near PCD7 = −0.4 creates a small "dip" before charges resume climbing. The general trend is positive: *areas with greater variance in median incomes (more income heterogeneity across the service area) tend to pay more.* The dip may reflect a specific regional cluster at that value.
- **PCD8** (Middle Income Dominated) Monotonically increasing from approximately PCD8 = −0.7 onward, with a flat segment below −0.4. Range of effect: ~653 to ~665.5 (approximately +12.5 units). The flat lower tail followed by an accelerating increase is similar to PCD1 in form—a threshold-then-ramp pattern. *Areas with more middle-income character reach a higher charge level once that character becomes sufficiently pronounced.*
- **ownership\_L** (public) contributes modestly positive SHAP in some utilities, suggesting ownership type modifies charges at the margin.

## Two-way Interactions

The interaction chart ranks overall feature interaction strength:

| Rank | Feature   | Interpretation                    |
|------|-----------|-----------------------------------|
| 1    | PCD2      | Wealthier, Urban Density & Rental |
| 2    | PCD1      | Economic Affluence                |
| 3    | PCD3      | Homeownership Strength            |
| 4    | ownership | Utility ownership type            |
| 5    | PCD8      | Middle Income Dominated           |
| 6    | PCD5      | Demographic Homogeneity           |

The fact that PCD1, PCD2, and PCD3 dominate **both** the importance ranking and the interaction ranking is significant: these are not merely additive drivers. Their effects are **context-dependent**, meaning the influence of Economic Affluence on charges differs depending on the urban character or homeownership profile of a service area. Ownership type is a meaningful moderator despite being lower in marginal importance—it amplifies or dampens demographic effects. We examine the nature of the interactions.

**PCD1 × PCD2:** The combined effect of Economic Affluence and Wealthier, Urban Density & Rental reveals that **neither dimension alone maximizes charges**—both must be elevated simultaneously.

- When PCD1 is negative (low economic affluence), charges remain low (~624–645) regardless of PCD2 level, except at the most extreme top-right corner where PCD2 is also very high (charges reach ~731).
- When PCD1 is positive but PCD2 is low (lower-left quadrant shifted right), charges remain at a moderate level (~667).
- The **maximum charges (~731)** occur at high PCD1 × high PCD2: economically affluent AND urban/wealthy service areas.
- A notable feature is an irregular boundary near  $PCD1 \approx -2$ ,  $PCD2 \approx 1-2$ , where a small pocket of elevated response exists—possibly a cluster of specific utility systems with these characteristics.

**Implication:** The PCD1 × PCD2 interaction suggests a **multiplicative premium** for areas that combine Economic Affluence with urban wealth and diversity. Infrastructure investment, service quality expectations, and cost recovery models likely all converge to produce this compounding effect.

**PCD1 × PCD4:** The PCD4 (Deep Poverty Gradient) axis shows minimal modulating effect on the primary PCD1 relationship.

- Charges increase dramatically with PCD1 across the entire range of PCD4 values—the horizontal color gradient confirms PCD1 dominates.
- At the most negative PCD1 and high PCD4 (upper-left), charges actually dip further (~645), suggesting that deep poverty compounds with economic instability to produce the lowest observed charges.
- The PCD4 axis contributes a modest vertical gradient: higher PCD4 is weakly associated with lower charges across all PCD1 levels.
- **Maximum charges (~691)** occur at high PCD1, with PCD4 near neutral—not at extreme poverty values.

**Implication:** Deep poverty (PCD4) slightly suppresses charges but plays second fiddle to the overriding effect of PCD1. The interaction is weak, consistent with PCD4's low Friedman H-statistic.

**PCD2 × PCD3:** Homeownership Strength (PCD3) meaningfully conditions how Wealthier, Urban Density & Rental (PCD2) translates into charges.

- The overall gradient runs diagonally: high PCD2 + positive PCD3 → higher charges (688); low PCD2 + negative PCD3 → lower charges (608).
- However, the response is not symmetric: at high PCD2 (right side), charges plateau regardless of PCD3 level—urban wealth saturates the response.
- At moderate PCD2 (~0 to +1), PCD3 becomes more important: the transition from negative to positive PCD3 shifts charges from ~657 to ~673.
- **The lowest charges (~608)** occur at the combination of **low PCD2 and very negative PCD3**: renter-heavy, non-urban, non-wealthy areas.

**Implication:** Homeownership amplifies the charge premium of urban wealth at intermediate urbanization levels, but has diminishing influence at the extreme high end of urban wealth. This is consistent with suburban homeowner communities at the margin of the urban core facing the steepest charge-homeownership gradient.

#### Geographic (State) Effects

State indicator variables for California, Texas, Washington, Pennsylvania, Mississippi, Oregon, Louisiana, New York, Georgia, and Alaska all appear in the top-20 feature importances and in the SHAP beeswarm plot.

The **state SHAP choropleth map** reveals a clear geographic structure:

- **Strong positive SHAP (high charges):** California and Alaska dominate as the states whose utility membership consistently elevates predicted charges above the national average, with mean SHAP values exceeding +60 for California utilities. Washington and Texas show moderate positive effects.
- **Strongly negative SHAP (low charges):** Mississippi stands out dramatically with a mean SHAP value approaching –60—the strongest negative state effect in the model. This implies that Mississippi utilities charge substantially below what their demographic characteristics alone would predict, suggesting structural differences in regulatory environment, cost structures, or subsidization.
- **Near-neutral states:** Most of the interior U.S. shows near-zero mean SHAP, indicating their charges are well-explained by demographic variables without requiring state-specific correction.

The persistence of strong state effects even after controlling for demographic PCs indicates that **regulatory frameworks, state-level policy environments, and infrastructure investment patterns** create charge premiums or discounts that are not fully captured by the demographic structure of service areas.

#### E1.1.1 Synthesis: Conceptual Model of Water Charge Drivers

Integrating all evidence, the following hierarchy and interaction structure emerges:

##### Tier 1: Primary Continuous Drivers (Socioeconomic Character)

**PCD2 (Wealthier, Urban Density & Rental )** is the single strongest demographic predictor, with a saturating positive effect. **PCD1 (Economic Affluence)** provides a complementary monotonic increase with a threshold near neutrality. Together, they account for the majority of charge variation attributable to service area demographics. Their interaction further amplifies charges in areas that combine both attributes.

## Tier 2: Geographic Regulatory Premium/Discount

State membership creates substantial, demographically-unexplained charge variation. California commands the largest positive premium; Mississippi the largest discount. These effects likely encode differences in regulatory philosophy, utility consolidation, infrastructure age, and cost-of-living-adjusted rate structures.

## Tier 3: Secondary Demographic Modifiers

**PCD3 (Homeownership Strength)**, **PCD5 (Demographic Homogeneity)**, **PCD7 (Median Income Variance)**, and **PCD8 (Middle Income Dominated)** each contribute moderate additive effects, with PCD3 operating most strongly as an interaction modifier of PCD2. The fact that PCD3 shows the third-highest Friedman H-statistic (behind only PCD2 and PCD1) confirms its role as a contextual amplifier.

## Tier 4: Complex/Interaction-Dependent Features

**PCD4 (Deep Poverty Gradient)** and **PCD6 (Income Inequality)** have small marginal effects with complex, non-monotonic shapes—diagnostic of features whose effects are primarily expressed through interactions with higher-tier variables. Their inclusion still improves model accuracy because they sharpen conditional predictions.

## Tier 5: Sparse/Threshold Effects

**PCS1 and PCS2** (service/seasonal components) show extreme spike-and-plateau PDPs concentrated in sparse data regions, suggesting their effects are specific to small subpopulations of utilities rather than broadly generalizable trends.

The prevalence of threshold and saturating relationships across the leading features has a practical implication: **linear models or simple additive rate benchmarking will systematically misestimate charges** for utilities whose demographic profiles lie near inflection points—particularly in the PCD1 = 0 to +2 and PCD2 = -0.5 to +1 transition zones.

### E1.1.2 Conclusions

1. **Socioeconomic structure is the dominant predictive signal.** Two dimensions—Economic Affluence and urban wealth/diversity—jointly explain most of the systematic variation in water charges across U.S. utilities. Any benchmarking or rate-adequacy analysis should explicitly control for these dimensions.
2. **Nonlinearity is the rule, not the exception.** The threshold effect in PCD1, the saturation in PCD2, and the inverted-U in PCD3 all indicate that charge effects are regime-dependent. Analyses assuming proportional effects of demographic variables will produce biased conclusions.
3. **State effects are large and demographically irreducible.** California's +60 SHAP and Mississippi's -60 SHAP are not explained by demographic composition alone. Regulatory, historical, and policy factors at the state level represent a major source of charge variation that must be addressed through policy analysis rather than demographic modeling.
4. **Interaction effects matter most for PCD1, PCD2, and PCD3.** Rate analysis for utilities in areas combining Economic Affluence with urban diversity should account for the superadditive interaction premium. Homeownership structure modifies how urban wealth translates to charges at intermediate urbanization levels.
5. **Surprisingly population or service size (PCS1 and PCS2) do not emerge as dominant factors** once the other demographic factors are accounted for.

## E1.2 10-year Health Violations

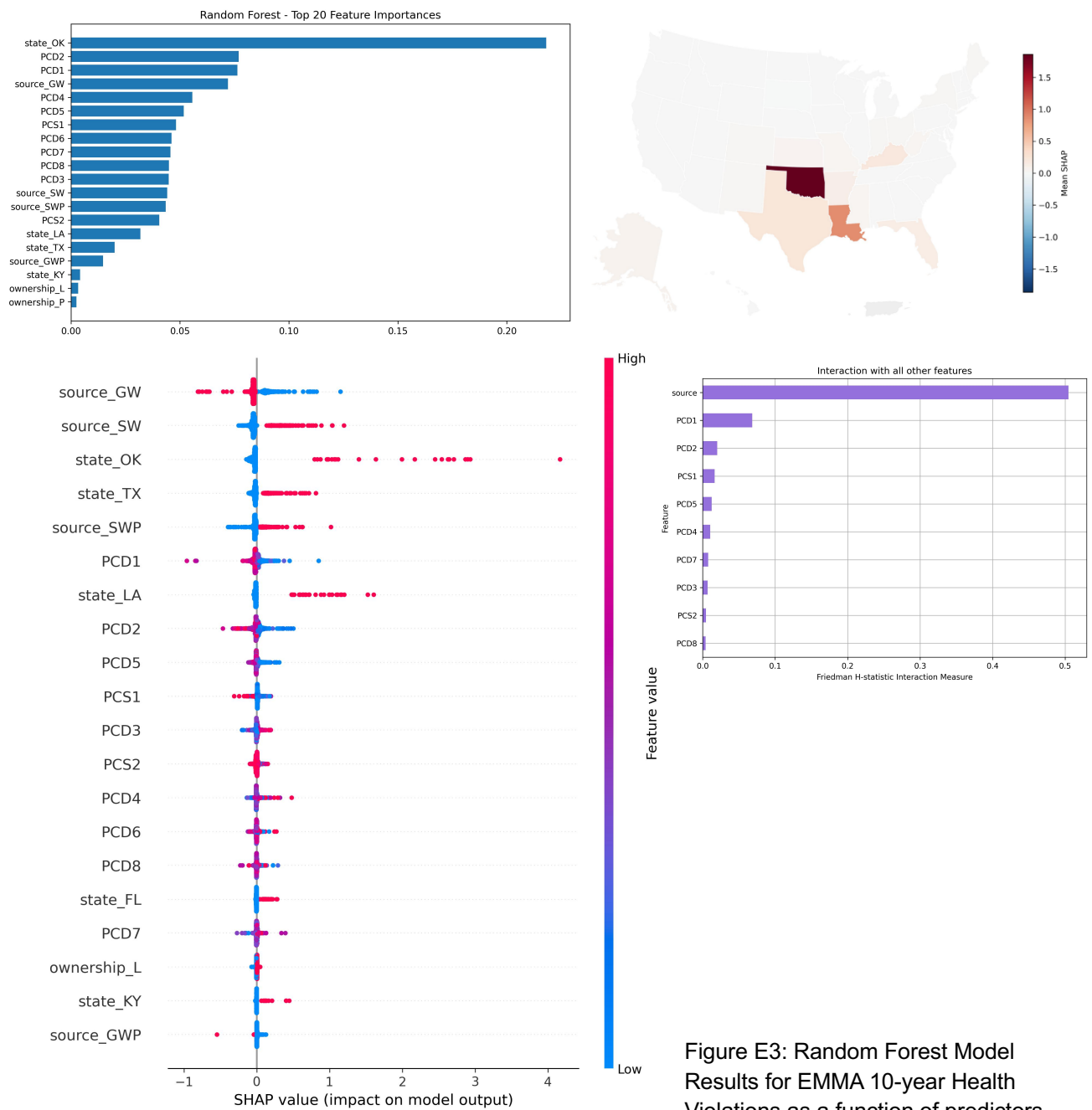


Figure E3: Random Forest Model Results for EMMA 10-year Health Violations as a function of predictors

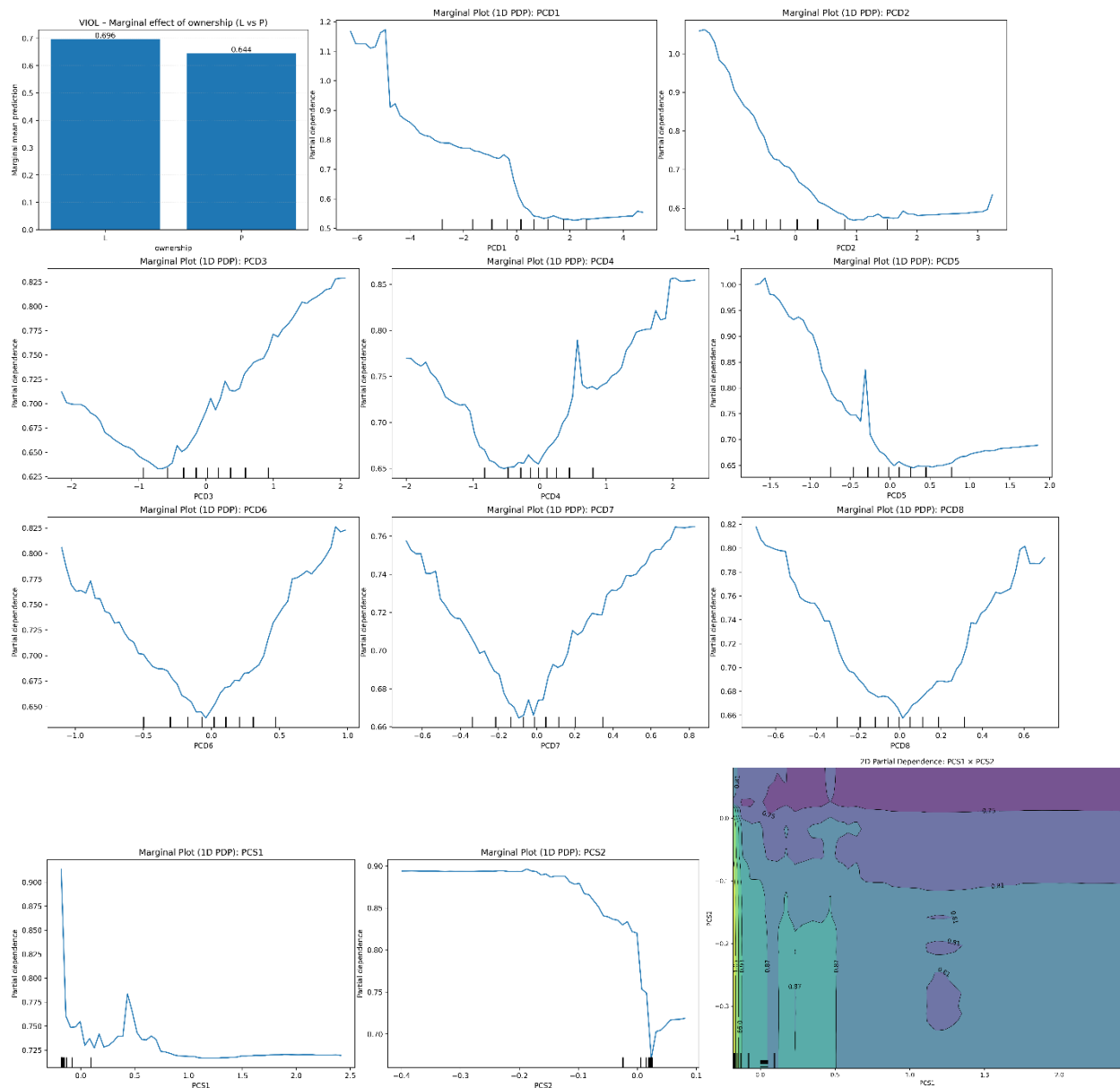


Figure E4: Random Forest Model Results for EMMA 10-year Health Violations for partial dependence of key predictors

The violation model tells a fundamentally different story from the charges model. **Water source type and state identity dominate**, with Oklahoma emerging as a singular outlier state. **The relationship between socioeconomic character and violations is largely inverse to charges**—economically stable, urban, wealthy service areas tend to have *fewer* violations, while system scale and service density play important moderating roles through the PCS components. Critically, violations exhibit pronounced nonlinearity: threshold effects, U-shaped responses, and extreme spike patterns all appear, indicating that the violation risk landscape is highly heterogeneous across the utility population.

### Feature Importance

The top-ranked features by mean decrease in impurity are:

| Rank | Feature    | Interpretation                                         |
|------|------------|--------------------------------------------------------|
| 1    | state_OK   | Oklahoma indicator                                     |
| 2    | PCD2       | Wealthier, Urban Density & Rental                      |
| 3    | PCD1       | Economic Affluence                                     |
| 4    | source_GW  | Groundwater source                                     |
| 5    | PCD4       | Deep Poverty Gradient                                  |
| 6    | PCD5       | Demographic Homogeneity                                |
| 7    | PCS1       | System Scale (avg. population served & connections)    |
| 8    | PCD6       | Income Inequality                                      |
| 9    | PCD7       | Median Income Variance                                 |
| 10   | PCD8       | Middle Income Dominated                                |
| 11   | PCD3       | Homeownership Strength                                 |
| 12   | source_SW  | Surface water source                                   |
| 13   | source_SWP | Surface water purchased                                |
| 14   | PCS2       | Service Density (contrast of population & connections) |
| 15   | state_LA   | Louisiana indicator                                    |
| 16   | state_TX   | Texas indicator                                        |

The most striking contrast with the charges model is that **state\_OK is now the single most important feature by a wide margin** (~0.22 MDI versus ~0.10 for PCD2), and **water source variables (GW, SW, SWP) now appear prominently** in the top features—they were essentially absent in the charges model. Ownership type (L and P) drops to near the bottom of the top-20 list. These structural shifts signal that violation risk is driven by fundamentally different mechanisms than pricing or ownership.

The SHAP beeswarm plot reveals the directional structure of each feature's contribution:

- **source\_GW** shows the most striking directional pattern: groundwater source utilities dominate the extreme right tail of positive SHAP values, with several cases exceeding +4. This is the highest magnitude individual-case SHAP effect in the model—certain groundwater utilities have violation predictions dramatically elevated above baseline. At low source\_GW values (surface water), SHAP values cluster near zero or slightly negative, confirming that the violation risk premium is specific to groundwater systems.

- **source\_SW** shows a complementary pattern: high surface water values produce modestly positive SHAP, but the effect is much weaker than groundwater. The surface water distribution is also right-skewed, suggesting a subset of surface water utilities with elevated violations.
- **state\_OK** produces a diffuse right-skewed positive SHAP distribution spanning 0 to beyond +4, consistent with its top importance ranking. Oklahoma utilities are systematically predicted to have more violations than their demographic and source characteristics alone would imply.
- **PCD1** (Economic Affluence) now shows **negative** SHAP at high values—the opposite direction from the charges model. Economically stable service areas predict *fewer* violations, consistent with better infrastructure investment and compliance capacity.
- **PCD2** (Wealthier, Urban Density & Rental ) similarly shows strongly negative SHAP at high values: urban, wealthy, diverse communities are associated with significantly lower violation counts. Combined with its charges-model behavior (high charges), this implies urban wealth buys both higher prices AND better compliance.
- **PCD5** (Demographic Homogeneity) exhibits a bimodal SHAP distribution with a substantial positive tail—high homogeneity is associated with more violations—consistent with its charges-model direction (higher charges) but here connecting to compliance risk rather than pricing.
- **state\_LA** (Louisiana) shows a clear rightward positive SHAP cluster, marking it as a secondary state-level violation risk factor, though far weaker than Oklahoma.
- **ownership\_L and ownership\_P** both have near-zero SHAP distributions—ownership type has negligible marginal impact on violations once other features are controlled, unlike in the charges model where it modestly appeared. This is an interesting finding.

#### *Geographic SHAP Pattern*

The state SHAP choropleth is dominated by two states with strong positive mean SHAP: **Oklahoma** (deep red, mean SHAP  $\approx +1.5$  to +2) and **Louisiana** (moderate red, mean SHAP  $\approx +0.5$ ). All other visible states show near-zero or very faintly positive/negative mean SHAP. This concentration is much more geographically narrow than the charges model, where California, Texas, Washington, and Mississippi all showed distinct effects. The implication is that violation risk—beyond what demographics and source type predict—is concentrated in specific regulatory and geological environments, with Oklahoma representing an extreme case likely linked to groundwater contamination exposure, agricultural activity, and the density of small, undercapitalized systems.

#### *Interaction Strengths (Friedman H-Statistic)*

| Rank | Feature | Interpretation                    |
|------|---------|-----------------------------------|
| 1    | source  | Water source type (dominant)      |
| 2    | PCD1    | Economic Affluence                |
| 3    | PCD2    | Wealthier, Urban Density & Rental |
| 4    | PCS1    | System Scale                      |
| 5    | PCD5    | Demographic Homogeneity           |
| 6    | PCD4    | Deep Poverty Gradient             |
| 7    | PCD7    | Median Income Variance            |
| 8    | PCD3    | Homeownership Strength            |

**Water source interaction strength (~0.5) is dramatically higher than all other features** and is roughly double the next-highest (PCD1 at ~0.2). This indicates that source type does not operate additively—it fundamentally changes how every other variable relates to violations. The effect of Economic Affluence, demographic structure, or system scale on violations is **conditioned on whether the system draws from groundwater versus surface water**. PCS1 (system scale) reaching the top-4 for interactions is also noteworthy: scale modulates how source type and demographics translate into violation risk, which the 2D PDP for PCS1×PCS2 illuminates directly.

**PCD1** (Economic Affluence) Monotonically decreasing with a steep drop near PCD1 = −5 to −4, a gradual decline through the negative range, then an accelerating drop between PCD1 = −1 and +0.5, followed by a flat low plateau. Range of effect: ~1.17 to ~0.53 (approximately −0.64 units; a ~55% reduction in predicted violations from extreme low to high PCD1).

*Nonlinearity:* Two-phase decline—a modest slope through economically unstable territory and then a sharp inflection near PCD1 = 0, where violations drop steeply before leveling off. This is the **mirror image of the charges model** threshold: where charges began accelerating upward at PCD1 ≈ 0, violations accelerate downward at essentially the same point. Economic Affluence is strongly protective against violations, but the protection mechanism kicks in most forcefully once a critical stability threshold is crossed. Utilities serving economically unstable populations maintain roughly similar (high) violation rates across the negative PCD1 range—there is no graduation of risk within depressed-economy territory, only a cliff at the transition to stability. This pattern suggests that compliance capacity is qualitatively different, not just quantitatively different, above versus below the stability threshold.

**PCD2** (Wealthier, Urban Density & Rental ) Monotonically decreasing throughout, with the steepest decline between PCD2 ≈ −1.8 and +0.8, then a flat low plateau, with a slight uptick at the extreme high end (PCD2 > 3). Range of effect: ~1.06 to ~0.57 (approximately −0.49 units; ~47% reduction). *Nonlinearity:* A sigmoidal decay that mirrors the sigmoidal increase seen in the charges model—the same demographic axis that drives prices up drives violations down, through the same mid-range transition zone. The slight uptick at the extreme right tail warrants attention: the handful of most urban, highest-PCD2 utilities may experience a modest violation rebound, possibly reflecting the complexity of serving very large, diverse urban infrastructure systems or reflecting data sparsity at the tail. Urban diversity and wealth is the second strongest demographic protector against violations. The saturation behavior implies that most of the violation-reduction benefit of urban wealth is captured at moderate levels; the most extreme urban wealth provides diminishing additional protection. This is consistent with regulatory capacity, infrastructure quality, and professional management all scaling with urbanization through the middle of the range and reaching a ceiling at the high end.

**PCD3** (Homeownership Strength) U-shaped (convex): starts elevated (0.714) at very negative PCD3, decreases to a minimum near PCD3 ≈ −0.7 (0.635), then increases monotonically and steeply through positive values to a maximum of ~0.830 at PCD3 ≈ +2.

Range of effect: ~0.635 to ~0.830 (approximately +0.195 units across the positive range).

*Nonlinearity:* A clear U-shape, but asymmetric—the minimum is displaced toward negative values and the right arm is steeper than the left. This is the inverse of what was observed for PCD3 in the charges model (where an inverted-U peak occurred at moderately negative values). Here, the *minimum* violation risk occurs at moderately negative PCD3 (moderate renter emphasis), while strong homeownership is associated with elevated violation risk.

The highest violation risk at high homeownership values likely reflects that strongly owner-occupied communities are often lower-density rural or exurban areas where small systems, aging private wells, and limited regulatory oversight create compliance vulnerabilities. The U-shape's left arm (elevated violations at very renter-heavy areas) may reflect dense, lower-income urban neighborhoods with aging infrastructure. The minimum at moderate homeownership represents a sweet spot—neither extremely renter-dominated nor strongly owner-dominated—that coincides with more professionally managed mid-tier utility systems.

**PCD4** (Deep Poverty Gradient) U-shaped with a sharp, narrow spike near  $PCD4 \approx +0.35$  (0.79), a broad minimum at  $PCD4 \approx -0.7$  to 0 (0.65), elevated values at negative extremes (0.77), and a strong monotonic increase from the spike onward to the maximum (0.855) at  $PCD4 \approx +2$ . Range of effect:  $\sim 0.65$  to  $\sim 0.855$  (approximately +0.205 units—the largest marginal range among the PCD features in this model).

*Nonlinearity:* Complex—a clear U-shaped base with a superimposed sharp spike and a strongly nonlinear right tail. The spike at  $PCD4 \approx +0.35$  is the most structurally interesting feature: it suggests a specific subset of utilities characterized by this particular poverty gradient value that have anomalously high violation rates. This is likely a small subgroup of utilities (consistent with the tick mark density in that region) where deep poverty combines with another factor to produce extreme violations. The strong positive slope at high PCD4 values is the clearest main effect: the deepest poverty gradients are the strongest demographic predictor of elevated violations in this model. Unlike in the charges model where deep poverty had a negligible marginal effect (operating mainly through interactions), here PCD4 has the largest marginal range of any demographic feature. Areas with severe poverty gradients face the highest violation risk, consistent with limited household ability to fund system upgrades, higher rates of small and undercapitalized systems, and reduced political capacity to pressure utilities for compliance. The spike at moderate PCD4 likely marks a specific utility subtype—possibly small rural systems at the intersection of poverty and specific source characteristics—that deserves targeted regulatory attention.

**PCD5** (Demographic Homogeneity) Strongly decreasing from a high (1.01) at very negative PCD5, with a sharp spike near  $PCD5 \approx -0.3$  (0.83) interrupting the general decline, then dropping to a flat low plateau ( $\sim 0.65$ – $0.69$ ) above  $PCD5 = 0$ . Range of effect:  $\sim 0.65$  to  $\sim 1.01$  (approximately  $-0.36$  units; the largest range among all PCD features). *Nonlinearity:* A dominant downward trend is overlaid by a sharp spike at  $PCD5 \approx -0.3$ . The general pattern—higher violations at low homogeneity—is counterintuitive at first but makes sense given the composition of the demographic homogeneity axis: low PCD5 likely describes highly heterogeneous demographic communities (often urban or diverse rural), some of which face structural disadvantages. However, the *charges model* showed violations increase at high PCD5 (high homogeneity), suggesting the relationship has reversed here. The spike near  $-0.3$  again marks a specific subgroup—likely a cluster of demographically heterogeneous small utilities with elevated exposure. The complex shape of PCD5 in this model, relative to its simpler monotonic form in the charges model, is diagnostic of strong interaction effects with source type. Demographic homogeneity's effect on violations likely differs substantially between groundwater and surface water systems, consistent with PCD5's top-5 interaction ranking. The spike is a data-sparse artifact or subgroup effect warranting investigation.

**PCD6** (Income Inequality) U-shaped (convex), symmetric around a minimum near  $PCD6 \approx 0$  (0.643). Elevated violations at both extreme negative (0.806) and extreme positive ( $\sim 0.826$ ) PCD6 values. Range of effect:  $\sim 0.643$  to  $\sim 0.826$  (approximately +0.183 units).

*Nonlinearity:* A clear, smooth U-shape that is one of the most interpretively transparent patterns in this model. Both extremes of income inequality—either very low or very high—are associated with more violations. Intermediate income inequality corresponds to the lowest violation rates. At very high income inequality, the pattern likely reflects communities where infrastructure investment is chronically underfunded because political power is concentrated among high-income households who can substitute private solutions (bottled water, filtration) for system compliance, while at low inequality, extreme uniformity may correlate with very small, homogeneous rural or isolated systems with limited capacity. The U-shape here is

a genuine and substantively interpretable relationship rather than a noise artifact, given the smooth curvature.

**PCD7** (Median Income Variance) V-shaped (convex), with a well-defined minimum near  $PCD7 \approx -0.05$  ( $\sim 0.664$ ) and symmetric elevation on both sides, reaching  $\sim 0.758$  at negative extremes and  $\sim 0.765$  at positive extremes. Range of effect:  $\sim 0.664$  to  $\sim 0.765$  (approximately  $+0.101$  units). *Nonlinearity*: A near-symmetric V-shape centered at  $PCD7 \approx 0$  is structurally similar to PCD6 but tighter and more symmetric, suggesting violations are minimized when median income shows average variance—neither very uniform nor very heterogeneous across the service area. Systems serving areas of average income uniformity (PCD7 near zero) have the fewest violations. Both very homogeneous income areas and very heterogeneous income areas show elevated risk. This likely reflects that extreme uniformity at low income equals concentrated poverty (high risk), extreme uniformity at high income equals wealthy areas prone to small private or community systems, and extreme heterogeneity captures mixed areas where infrastructure planning is complicated.

**PCD8** (Middle Income Dominated) U-shaped with a sharp, well-defined minimum near  $PCD8 \approx +0.05$  ( $0.660$ ). Strongly elevated at negative extremes ( $0.820$ ) and elevated at positive extremes ( $\sim 0.800$ ). Range of effect:  $\sim 0.660$  to  $\sim 0.820$  (approximately  $+0.160$  units).

*Nonlinearity*: Similar to PCD6 and PCD7 in overall form—convex U—but with the minimum displaced slightly into positive values. A striking feature is that both extremes of middle-income dominance are associated with higher violations, with the negative extreme (non-middle-income dominated, i.e., bipolar income distributions) showing slightly higher risk than the positive extreme. Middle-income-dominated areas minimize violation risk, a finding that aligns intuitively: middle-income communities are typically large enough to support professional utility management, have sufficient political voice to demand compliance, and serve populations with stable water infrastructure investment capacity. Extreme deviation in either direction—toward bimodal poverty/wealth or toward very narrow class composition—both elevate risk.

**PCS1** (System Scale) Spike-and-decay with a high peak near  $PCS1 \approx -0.1$  ( $0.915$ ), rapid decay through  $PCS1 = 0$  to  $\sim 0.75$ , a secondary spike at  $PCS1 \approx +0.4$  ( $0.783$ ), then gradual decline to a stable low plateau ( $\sim 0.718$ ) for  $PCS1 > 0.7$ . Range of effect:  $\sim 0.718$  to  $\sim 0.915$  (approximately  $-0.197$  units from smallest to largest systems). *Nonlinearity*: Unlike the charges model where PCS1 showed a single sharp spike at the left boundary, here the decay is more gradual after the initial peak, with a secondary feature at  $PCS1 \approx +0.4$ . The plateau behavior at large system scale is clear: large utilities converge to a stable, lower violation rate. The secondary spike warrants scrutiny—it likely marks a specific cluster of mid-scale systems, possibly in specific states (Oklahoma or Louisiana), that interrupt the general scale-protection trend. Small systems (low PCS1) have dramatically higher violation rates than large ones, consistent with extensive prior research on system scale and regulatory compliance. Larger systems benefit from economies of scale in monitoring, treatment, and regulatory engagement. The flat plateau above  $PCS1 \approx 0.7$  suggests that once systems reach a meaningful minimum scale, additional scale provides diminishing additional protection. The secondary spike interrupts this narrative and likely reflects a scale-state or scale-source interaction rather than a genuine non-monotonic scale effect.

**PCS2** (Service Density) Flat high plateau ( $0.895$ ) for all  $PCS2 < -0.05$ , then a sharp cliff drop between  $PCS2 \approx -0.05$  and  $+0.02$ , reaching a minimum ( $0.690$ ) near  $PCS2 \approx +0.02$ , followed by a partial recovery to  $\sim 0.717$  at the right boundary. Range of effect:  $\sim 0.690$  to  $\sim 0.895$  (approximately  $-0.205$  units—one of the largest marginal ranges in this model). *Nonlinearity*: An almost pure threshold/step-function: violation risk is uniformly high across essentially all negative PCS2 values, then drops sharply at the zero crossing. This is one of the most structurally clean nonlinearities in the model. The tick marks confirm the data is concentrated near the threshold region. PCS2 represents the contrast between population served and service connections—a proxy for service density. Negative PCS2 indicates more connections relative to population (low connection density, consistent with dispersed rural systems or systems with many small

customers per connection). Positive PCS2 indicates fewer connections relative to population (higher density, urban or suburban character). The sharp threshold at  $PCS2 \approx 0$  suggests that a fundamental shift in system character—from dispersed to dense service—is associated with a step-change in compliance. Dispersed, low-density systems have uniformly high violation rates regardless of exactly how dispersed they are; denser systems uniformly have lower rates. This reinforces the PCS1 finding that system organizational character, not just size, shapes compliance capacity.

**PCS1 × PCS2:** The joint PCS1 × PCS2 2D PDP reveals that **service density (PCS2) is the primary organizing dimension for violations, with system scale (PCS1) playing a secondary modulating role at small scales only.**

- The dominant feature of the plot is the horizontal banding: the PCS2 axis creates a sharp boundary between the high-violation upper region (PCS2 near zero and positive, purple/blue band at top) and the lower-violation middle region (PCS2 negative, teal). This corresponds directly to the PCS2 1D PDP threshold.
- At very small system scale ( $PCS1 \approx 0$ ) and negative PCS2, violations reach their maximum ( $\sim 1.05$ ). This left-edge column of yellow-green marks the most dangerous system archetype: small AND low-density.
- As PCS1 increases beyond 0.5, ~~violations decline to a stable teal plateau (0.81)~~ that is largely independent of PCS2 for the PCS2-negative region. Scale appears to provide protection primarily for the smallest systems; once past a threshold, additional scale does not further reduce violations.
- The upper band (PCS2 near zero) maintains elevated violations ( $\sim 0.75$ – $0.81$ ) even at large PCS1, confirming that service density is a binding constraint that scale cannot overcome.
- Isolated elevated patches at  $PCS1 \approx 0.1$ – $0.5$ ,  $PCS2 \approx -0.05$  to  $-0.1$  (contour values  $\sim 0.87$ ) represent a specific small-to-medium-scale, moderately-low-density system archetype with elevated violation risk—likely corresponding to particular state/source combinations.

**Implication:** The violation-reduction benefits of system scale are captured primarily in the transition from very small ( $PCS1 < 0.2$ ) to small-medium ( $PCS1 \approx 0.2$ – $0.5$ ). Beyond that, service density dominates. Policies targeting scale consolidation of the smallest systems will have the greatest per-merger compliance benefit; additional consolidation above that threshold provides diminishing returns unless it also shifts systems to denser service configurations.

### *Ownership Effect*

The marginal effect of ownership is clear and directionally unambiguous: **locally-owned (L) utilities have a higher predicted violation rate (0.696) than privately-owned (P) utilities (0.644)**, a difference of approximately 0.052. This is a modest but consistent effect, after accounting for all other factors

This finding contrasts somewhat with the commonly assumed advantage of public ownership in terms of social mission. In the context of this model, locally-owned systems likely disproportionately represent the small, rural, low-density system archetypes identified above as highest-risk. The private ownership premium in compliance may reflect the concentration of large investor-owned utilities (which are typically larger, better-capitalized, and subject to rigorous investor and regulatory scrutiny) in the private ownership category. The effect is small relative to source type, state, and demographic drivers, consistent with ownership's near-bottom ranking in both importance and interaction statistics.

### *Water Source Effects*

Source type (groundwater vs. surface water) is the dominant **interaction variable** in the model (Friedman  $H \approx 0.50$ , approximately twice the next-highest feature). The SHAP beeswarm confirms that groundwater systems produce the most extreme positive SHAP outliers—individual utilities with groundwater sourcing

can be predicted to have violations far above baseline, while no surface-water system reaches comparable extremes.

The mechanism is well-established: groundwater systems face different (and in many rural contexts more heterogeneous) contamination risks than surface water systems. They are also more likely to be small and dispersed, compounding the PCS1/PCS2 risk factors. The extreme right-tail SHAP values for source\_GW likely correspond to small Oklahoma or Louisiana groundwater systems, explaining why the state × source combination drives the most extreme predictions.

Surface water purchased (source\_SWP) shows a smaller but directionally positive SHAP effect, consistent with the additional compliance complexity of purchased water systems (dependence on a wholesaler, potential monitoring gaps).

### *E1.2.1 Synthesis: Conceptual Model of Health Violation Drivers*

#### *Tier 1: State × Source Interaction (Contextual Override)*

**Oklahoma** and, to a lesser extent, **Louisiana** have violation rate premiums that are not fully explained by demographics or source type alone, and groundwater sourcing is the single strongest interaction variable in the model. The **combination of Oklahoma + groundwater is almost certainly the highest-violation archetype** in the data. These effects likely encode regulatory history, geological contamination exposure (nitrates, arsenic, agricultural runoff), and the density of small underfunded systems that characterize both states.

#### *Tier 2: System Scale and Service Density (Infrastructure Character)*

**PCS1** (scale) and **PCS2** (density) together define the infrastructure character of a utility. Their 2D interaction shows that the highest-risk archetype is small AND low-density, and that neither scale nor density protection alone is sufficient—both must cross threshold values for risk to drop substantially. These are the most directly actionable levers for policy (consolidation, infrastructure investment), as they do not require changing the demographic or geographic context of service.

#### *Tier 3: Socioeconomic Character (Inverse of Charges Model)*

**PCD1** (Economic Affluence) and **PCD2** (Wealthier, Urban Density & Rental ) are the strongest demographic drivers of violations and operate in the opposite direction from their charges effects: prosperity and urbanization protect against violations. **PCD4** (Deep Poverty Gradient) has the largest marginal range of any demographic feature and operates in the expected direction: extreme poverty maximizes violation risk. **PCD5** (Demographic Homogeneity) and the income structure dimensions (PCD6, PCD7, PCD8) all show U-shaped or convex responses—middle-ground demographic configurations minimize violations, with extremes in either direction elevating risk.

#### *Tier 4: Homeownership and Income Distribution Structure*

**PCD3** (Homeownership Strength) shows a U-shape where the minimum risk falls at moderate homeownership levels. High homeownership elevates risk (rural/exurban small systems), and very low homeownership also elevates risk (dense, poor urban infrastructure). PCD6, PCD7, and PCD8 similarly show convex forms, suggesting violations are minimized at typical (moderate variance, moderate inequality, middle-income-dominated) demographic profiles—and that extreme demographic configurations in any direction introduce compliance risk.

#### *Tier 5: Ownership (Marginal Modulator)*

**Ownership type** (L vs. P) has a small but consistent effect: privately-owned utilities average slightly fewer violations. This effect is likely mediated by the scale and source characteristics typical of each ownership category rather than ownership per se.

### Key Contrasts with the Water Charges Model

The violation and charges models share the same features but tell opposite stories for several key variables:

| Feature                   | Charges Direction                 | Violations Direction              | Interpretation                                                |
|---------------------------|-----------------------------------|-----------------------------------|---------------------------------------------------------------|
| PCD1 (Economic Affluence) | Positive monotone, threshold at 0 | Negative monotone, cliff at 0     | Affluence raises prices AND reduces violations                |
| PCD2 (Urban Wealth)       | Positive sigmoid, saturating      | Negative sigmoid, saturating      | Urban wealth → higher charges AND better compliance           |
| PCD3 (Homeownership)      | Inverted-U, peak at low values    | U-shape, minimum at low values    | Optimal ownership regime is similar but relationship inverted |
| PCS1 (Scale)              | Spike artifact, near-flat         | Protective decay, spike + plateau | Scale matters for violations far more than charges            |
| PCS2 (Density)            | Spike artifact                    | Sharp step-function protector     | Density is a major violation predictor, not a charge driver   |
| Source type               | Absent from top features          | Dominant interaction variable     | Source drives violations, not charges                         |
| State effects             | Broad (CA, TX, WA, MS, AK)        | Concentrated (OK, LA)             | Violation geography is more geographically concentrated       |

The mirror-image behavior of PCD1 and PCD2 across the two models is particularly striking: the same demographic axis that explains why wealthy, stable urban areas pay more also explains why they have fewer violations. This suggests that **socioeconomic structure does not create a trade-off between affordability and safety**, prosperity delivers both higher charges and better compliance simultaneously, while economic disadvantage produces both lower charges and greater violation risk.

### Key Nonlinearities and Their Substantive Implications

| Feature          | Nonlinear Pattern                     | Implication                                                                       |
|------------------|---------------------------------------|-----------------------------------------------------------------------------------|
| PCD1             | Cliff at 0, flat on both sides        | Compliance capacity threshold; crossing it matters far more than degree           |
| PCD2             | Saturating sigmoid decay              | Moderate urbanization provides most of the protection; extreme wealth adds little |
| PCD4             | U-base + spike + strong positive tail | Extreme poverty is the highest-risk demographic configuration                     |
| PCD6, PCD7, PCD8 | U-shaped / V-shaped                   | Middle-ground income structures minimize violations; extremes elevate risk        |
| PCS2             | Step-function at ~0                   | Service density is a binary-like threshold, not a continuous protector            |

| Feature               | Nonlinear Pattern                                       | Implication                                                          |
|-----------------------|---------------------------------------------------------|----------------------------------------------------------------------|
| PCS1 × PCS2           | Multiplicative: smallest AND least dense = maximum risk | Neither scale nor density alone is sufficient; both must be adequate |
| source × demographics | Dominant interaction ( $H \approx 0.5$ )                | All demographic effects are conditioned on source type               |

### E1.2.2 Conclusions

**1. Groundwater systems in Oklahoma and Louisiana represent a concentrated, identifiable high-violation population.** The combination of top-ranked state and top-ranked interaction variable points to a specific regulatory intervention target. State-level groundwater monitoring programs in these states warrant priority resourcing.

**2. System scale and service density consolidation offers the clearest policy leverage.** The PCS1 × PCS2 interaction identifies the smallest, most dispersed systems as the highest-risk archetype. Consolidation programs that bring sub-threshold systems above the PCS1  $\approx 0.2$ – $0.5$  and PCS2  $\approx 0$  thresholds could produce the largest per-action violation reduction.

**3. Socioeconomic disadvantage compounds violation risk nonlinearly.** The PCD1 cliff and PCD4 strong right tail both indicate that the most economically distressed communities face qualitatively—not just quantitatively—higher violation risk. Compliance assistance programs should be targeted to the tail of the economic disadvantage distribution, not uniformly distributed.

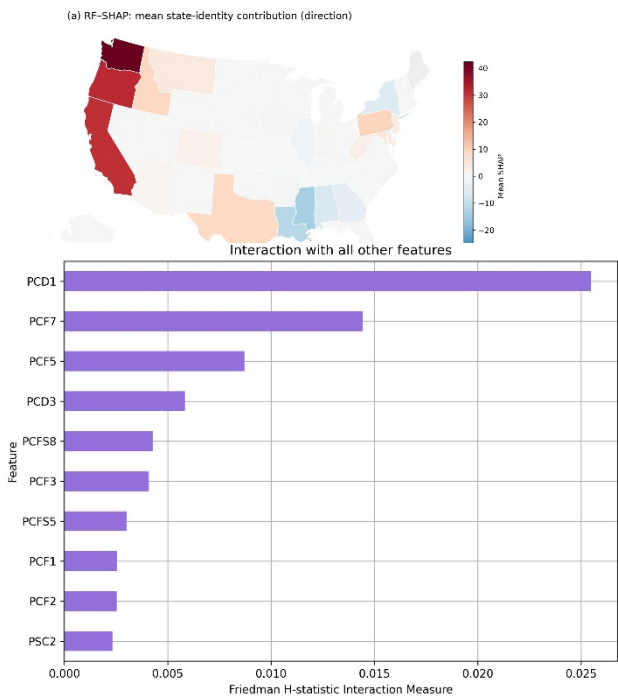
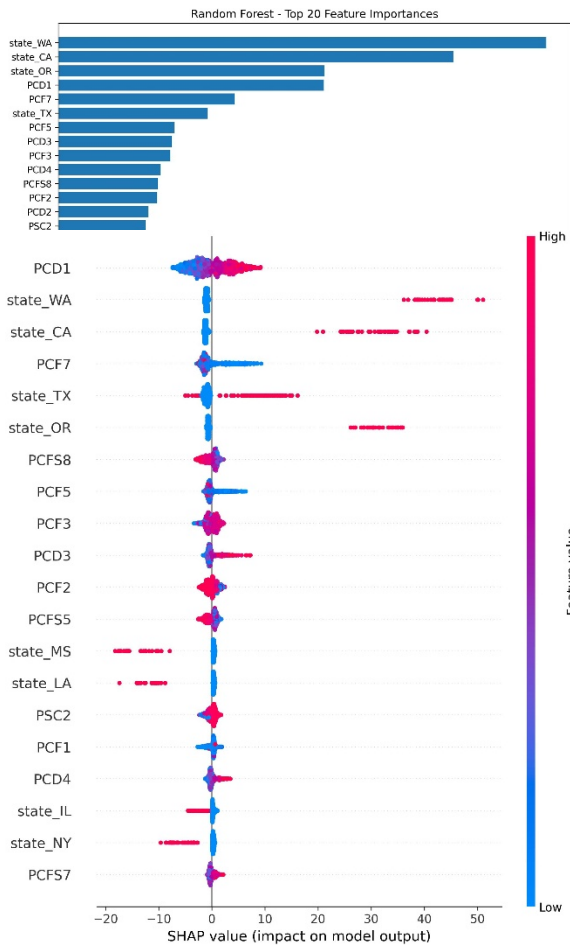
**4. The U-shaped responses of PCD6, PCD7, and PCD8 suggest that income structure extremes are risk factors regardless of direction.** Both very equal and very unequal communities face elevated violations compared to average-inequality communities. Policy interventions should account for this heterogeneity rather than treating income inequality as a simple linear risk factor.

**5. Ownership type has a small but real effect.** The 7.5% higher marginal violation rate for locally-owned utilities, while modest, is consistent across the model. Targeted technical assistance to small public utilities—particularly in high-risk state/source combinations—is warranted.

**6. The mirror relationship between charges and violations has affordability implications.** Because prosperity both raises charges and reduces violations, low-income communities face a compound disadvantage: they pay less in absolute terms but receive worse compliance. Regulatory frameworks that use low average charges as evidence of adequate service may be systematically underserving the highest-risk populations.

E2. Public

E2.1Average Water Charges



| train_R2 | cv_R2 |
|----------|-------|
| 0.52     | 0.32  |

Figure E5 Random Forest Model Results for Average Water Charges of Public Water Utilities

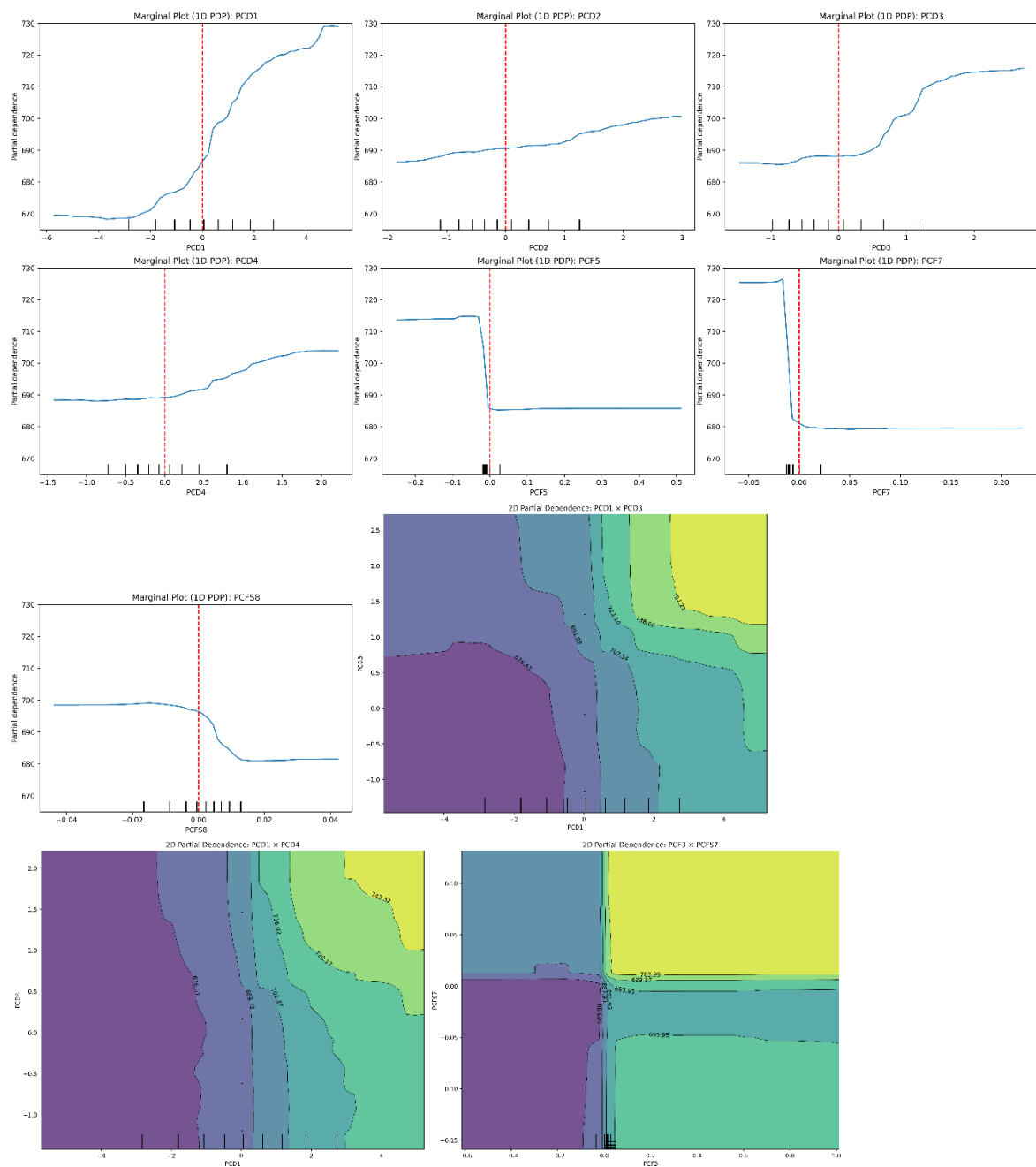


Figure E6 Partial Dependence Plots for Random Forest Model Results for Average Water Charges of Public Water Utilities

A Random Forest (RF) model was trained to predict average water charges for **public water utilities** across the United States. The predictor set combines state indicators, demographic principal components describing socioeconomic vulnerability (PCD1–PCD8), aggregate financial principal components (PCF1–PCF10), per-capita scaled financial principal components (PCFS1–PCFS10), and system-scale variable for population served and contrast between population and service connections (PSC1 and PSC2). The response variable is average annual water charge.

The public utility charges model occupies a distinctive position in the series of models analyzed. It most closely resembles the full-utility charges model in that **state identity is the dominant signal**—Washington,

California, and Oregon lead the importance rankings by wide margins—but it differs substantially in the nature of the demographic effects, which are now built around poverty and vulnerability gradients rather than affluence and urban sophistication. Critically, in this model **PCD1 operates as a continuously increasing function across its entire range** rather than a step function, sigmoid, or threshold ramp. The financial variables appear in secondary roles and show the sharp zero-crossing thresholds familiar from the private utility models, but with directions that reflect the public finance context: debt leverage (PCF7) and Revenue Expense Contrast (PCF5) are the leading financial predictors. The overall picture is one where **geography and poverty structure jointly determine public utility charges**, with financial characteristics providing context-dependent amplification, and the interaction structure confirming that poverty and health vulnerability combine multiplicatively to push charges higher in the most disadvantaged communities served by public utilities.

### Feature Importance

The top-ranked features are:

| Rank | Feature  | Category              | Interpretation                      |
|------|----------|-----------------------|-------------------------------------|
| 1    | state_WA | Geographic            | Washington state indicator          |
| 2    | state_CA | Geographic            | California state indicator          |
| 3    | state_OR | Geographic            | Oregon state indicator              |
| 4    | PCD1     | Demographic           | Economic Affluence                  |
| 5    | PCF7     | Financial (aggregate) | Debt Leverage                       |
| 6    | state_TX | Geographic            | Texas state indicator               |
| 7    | PCF5     | Financial (aggregate) | Revenue Expense Contrast            |
| 8    | PCD3     | Demographic           | Health Insurance Gap                |
| 9    | PCF3     | Financial (aggregate) | Taxation Profile                    |
| 10   | PCD4     | Demographic           | Renting Poor vs. Owning Poor        |
| 11   | PCFS8    | Financial (scaled)    | Utility Tax Contribution per capita |
| 12   | PCF2     | Financial (aggregate) | Revenue/Expense Volume              |
| 13   | PCD2     | Demographic           | Urban Mobility & Tenure             |
| 14   | PSC2     | System scale          | System scale component 2            |
| 15   | state_LA | Geographic            | Louisiana state indicator           |
| 16   | PCF1     | Financial (aggregate) | Overall Utility Scale               |
| 17   | PCD5     | Demographic           | Vehicle Dependency                  |
| 18   | state_MS | Geographic            | Mississippi state indicator         |
| 19   | PCD7     | Demographic           | Income Inequality Gap               |
| 20   | PCD6     | Demographic           | Low-Income Edge Cases               |

Three structural features stand out immediately. First, **three Western states (WA, OR, CA) occupy the top three positions**, with Washington having an MDI roughly double that of California. This West Coast dominance is new to this model—the full-utility charges model showed California dominant but not Washington or Oregon at this scale. Second, **PCD1 (Economic Affluence) ranks 4th overall**—the highest any poverty-focused demographic variable has ranked in any model, confirming that public utility charges in poverty-concentrated areas are genuinely elevated, a finding that inverts the typical assumption that poorer areas pay less. Third, **PCF7 (Debt Leverage)** at rank 5 introduces a long-term liability dimension absent from all prior models; public utility charges are partially set to service debt in a way that investor-owned utility rates are not structured to reflect directly.

The SHAP beeswarm reveals the directional story:

- **PCD1 (Economic Affluence):** Produces the most complex and substantively rich SHAP distribution in the model. High-value (red/pink) dots extend to the right (+5 to +10 SHAP), while low-value (blue) dots cluster near zero or slightly negative. The distribution is continuous and centered, consistent with PCD1's monotonically increasing PDP—there is no binary regime, but rather a graded positive relationship. **Higher poverty concentration in the service area predicts higher public utility charges, which is a counterintuitive finding that deserves emphasis.** The SHAP spread is moderate (+10 to −5 units), confirming PCD1 is a real but not extreme driver of charges.
- **state\_WA:** All positive SHAP values, clustered tightly between +35 and +52—Washington public utilities charge dramatically above the national average predicted by their demographic and financial characteristics. The consistency and magnitude (essentially no Washington utility is below the premium) suggests a state-wide pricing environment rather than a subgroup effect.
- **state\_CA:** Similarly all positive, ranging +20 to +40. California's public utility premium is slightly smaller in magnitude than Washington's but equally consistent across utilities.
- **state\_OR:** All positive, ranging roughly +20 to +35. Oregon forms a coherent West Coast high-charge cluster with Washington and California.
- **state\_TX:** A notable bimodal pattern—many red (high-value) dots generate modestly positive SHAP (+5 to +17), while a mass of blue dots cluster near zero. Texas public utilities are split: some charge well above baseline, others are average. The mixed pattern likely reflects Texas's highly heterogeneous public utility landscape (very large metropolitan utilities vs. small rural districts).
- **PCF7 (Debt Leverage):** Blue (low-value) dots extend to the right (+3 to +8 SHAP)—counterintuitively, utilities with *low* long-term debt relative to revenue (negative PCF7) are associated with higher charges. High debt utilities show near-zero or slightly negative SHAP. The Partial Dependence Plot will further illuminate this.
- **PCFS8 (Utility Tax Contribution per capita):** Mixed SHAP centered near zero, with red dots slightly left and blue dots slightly right—higher per-capita tax contribution is marginally associated with lower charges, and lower tax contribution with higher charges. This suggests utilities contributing more to general tax revenue are either large-scale (spreading costs) or operating in contexts where fiscal capacity allows lower rates.
- **PCF5 (Revenue Expense Contrast):** Blue dots dominate the right side—low Revenue Expense Contrast (negative PCF5, where expenses exceed revenues) generates positive SHAP, meaning utilities with poor revenue capture charge more. This is directionally consistent with a cost-recovery pricing model.
- **state\_MS and state\_LA:** Both show entirely leftward SHAP distributions (negative values, ranging −5 to −20), confirming these states' public utilities charge substantially below what their demographic and financial profiles would predict—consistent with the full-utility model finding of Mississippi's strong negative SHAP.

- **state\_NY and state\_IL:** Both show modestly negative SHAP, suggesting public utilities in these states charge somewhat below model-predicted levels, possibly reflecting municipal rate-setting constraints in older, dense urban environments.
- **PCD3 (Health Insurance Gap):** High values generate modestly positive SHAP (+2 to +8), while low values sit near zero. Communities with more uninsured residents—financially vulnerable populations—are associated with higher public utility charges. This reinforces the broader PCD1 finding and is again counterintuitive at first glance.

The state SHAP maps provide a striking geographic picture. The direction map shows a clear **West Coast premium** (California, Oregon, Washington all deep red, mean SHAP +20 to +45) alongside a clear **Deep South discount** (Mississippi, Louisiana deep blue, mean SHAP –20 to –30). Texas shows a moderate positive effect (+5 to +15). Pennsylvania appears modestly positive, consistent with its elevated private utility effects. The magnitude map confirms California as the largest-magnitude state effect (black), with Washington and Oregon also in the highest tier. Mississippi and Louisiana are second-tier in magnitude (dark purple).

This geographic structure has a coherent interpretation. The West Coast premium reflects the combination of California's historically high utility rates (driven by water scarcity, infrastructure investment, and environmental compliance costs), Washington's hydropower-adjacent public utility sector with specific rate structures, and Oregon's alignment with this regional pricing environment. The Deep South discount reflects persistently low public utility rates in Mississippi and Louisiana, likely tied to lower median incomes, lower infrastructure investment expectations, and historically less active utility rate reform—consistent with findings from all prior models in this series.

#### Interaction Strengths (Friedman H-Statistic)

| Rank | Feature | H-Statistic (approx.)            |
|------|---------|----------------------------------|
| 1    | PCD1    | ~0.026 (dominant by wide margin) |
| 2    | PCF7    | ~0.015                           |
| 3    | PCF5    | ~0.009                           |
| 4    | PCD3    | ~0.006                           |
| 5    | PCFS8   | ~0.005                           |
| 6    | PCF3    | ~0.005                           |
| 7    | PCFS5   | ~0.003                           |
| 8    | PCF1    | ~0.002                           |
| 9    | PCF2    | ~0.002                           |
| 10   | PSC2    | ~0.002                           |

**PCD1's H-statistic (~0.026) is the largest interaction measure of any single feature across all four models analyzed in this series**—roughly 70% larger than the next highest (PCF7 at ~0.015) and orders of magnitude larger than the H-statistics seen in the private utility models. This means PCD1 is not only the leading continuous demographic predictor but also the primary interaction variable: the Economic Affluence fundamentally changes how every other feature—financial structure, state, system scale—translates into charges. PCF7 at rank 2 is also a strong interaction variable, and its pairing with PCF5 and PCFS8 in the interaction rankings confirms that the public finance dimension (debt leverage, Revenue Expense Contrast, per-capita tax contribution) creates conditional charge effects that depend heavily on the demographic and geographic context of the utility. PCD3's rank-4 interaction position despite its rank-8 importance confirms that the Health Insurance Gap is primarily an interaction modifier rather than a direct charge driver.

### Partial Dependence Analysis

**PCD1 (Economic Affluence)** A continuously increasing, near-monotone curve spanning the entire support range (approximately -6 to +5). The curve rises from approximately 669 at the most negative values through a gentle, nearly linear slope across the negative domain, then accelerates into a steeper rise above PCD1 = 0, reaching approximately 729 at the most positive values. The red dashed line at zero marks a mild inflection point where the slope steepens noticeably. Range of effect: ~669 to ~729 (approximately +60 units; ~9% increase—the largest marginal range of any continuous feature in the public utility charges model). *Nonlinearity:* Modest but genuine—the slope steepens around zero, creating a slight concave-up (accelerating) form in the positive range. Public utilities serving areas with higher poverty concentration charge *more*, not less—a result that holds across the entire PCD1 range and accelerates above the mean. This paradox is the central demographic finding of the public utility charges model. The mechanism is likely cost-based: poverty-concentrated areas often have older infrastructure requiring more intensive maintenance and capital renewal, lower consumption per connection (reducing economies of scale in revenue recovery), and weaker local fiscal capacity to subsidize rates through general fund transfers. Public utilities serving wealthy areas may also benefit from more political pressure to keep rates low from influential ratepayers. The continuous, non-threshold form of this relationship suggests there is no poverty level at which the pricing premium activates discretely—it accumulates gradually across the spectrum.

**PCD2 (Urban Mobility & Tenure)** Slowly and nearly linearly increasing from approximately 686 at PCD2 = -2 to approximately 701 at PCD2 = +3. The red dashed line at zero falls in the middle of a smooth, near-linear progression with no inflection, plateau, or threshold. Range of effect: ~686 to ~701 (approximately +15 units; ~2.2%—a modest marginal range).

*Nonlinearity:* Minimal—this is the most nearly linear PDP across all features in all four models. The absence of threshold behavior is notable and consistent with PCD2's lower importance ranking (rank 13) and absence from the interaction chart top positions. Dense urban populations characterized by high rental rates and low vehicle ownership pay modestly more for public utility water. The linear, gradual form suggests urban mobility and rental tenure are contributing to charges through a simple demographic-cost relationship (higher density, older urban infrastructure, higher operating complexity) rather than through any regime-switching mechanism. The modest range confirms this is a background effect rather than a primary driver.

**PCD3 (Health Insurance Gap)** Flat to slightly declining from PCD3 ≈ -1.5 to approximately 0 (range 686–688), then a distinct acceleration beginning just above zero, with a steeper rise from PCD3 = 0 to +0.5 (climbing to ~700), followed by a further increase through a ~~secondary kink at PCD3 ≈ +1.1 (710)~~, reaching a plateau near ~716 for PCD3 > +2.

Range of effect: ~686 to ~716 (approximately +30 units; ~4.4%). *Nonlinearity:* A two-stage increase: nearly flat in the negative domain, then accelerating past zero with a secondary kink at moderate positive values before plateauing. The zero crossing is meaningful—uninsured rates at or below the mean produce essentially no charge premium, while above-average health insurance gaps generate a substantial and

accelerating charge premium. The plateau above  $PCD3 \approx +2$  suggests saturation: the most health-insecure communities are already at the maximum predicted charge level attributable to this variable. Public utilities serving communities with high rates of uninsured residents charge substantially more. This aligns with the PCD1 poverty finding: health insurance gaps and poverty are positively correlated, and both independently predict higher public charges. The threshold-like acceleration above zero suggests that utilities tolerate modest health vulnerability without charge consequences, but once uninsured rates exceed the community norm, a qualitatively different infrastructure or fiscal environment applies. The plateau at high PCD3 may reflect that communities with extreme health insurance gaps are already captured by the poverty gradient (PCD1) and the marginal contribution of PCD3 above that level is limited.

**PCD4 (Renting Poor vs. Owning Poor)** Essentially flat and near zero at approximately 689 for all  $PCD4 < 0$ , then a continuous and accelerating rise from  $PCD4 = 0$  through approximately +2.2, reaching ~704. The red dashed line marks the transition from flat to rising territory precisely. Range of effect: ~689 to ~704 (approximately +15 units; ~2.2%).

*Nonlinearity:* A classic threshold pattern: flat below zero, rising above. The rise is gradual and monotone rather than a sharp step, distinguishing it from the binary thresholds in the private utility models. The threshold at zero is meaningful: PCD4 contrasts impoverished renters (negative values) against impoverished non-renters/owners (positive values). The zero crossing represents parity between these two poverty subtypes. PCD4 distinguishes communities where poverty is concentrated among homeowners or stable housing residents (positive PCD4) from those where poverty is concentrated among renters (negative PCD4). Public utilities serving non-renter poor communities charge more than those serving renter poor communities. This is consistent with the spatial and infrastructure economics of public utilities: homeowner-poor areas (often rural, low-density, and with aged private infrastructure connecting to the public system) impose higher per-connection service costs, while renter-poor areas (often urban/dense) benefit from infrastructure density and scale. The flat left tail confirms that the renter-poor demographic does not create a charge premium relative to the overall mean.

**PCF5 (Revenue Expense Contrast)** A negative step-function: flat high plateau (714) for all  $PCF5 < -0.04$ , a steep drop between  $PCF5 \approx -0.04$  and 0, flat low plateau (686) for all  $PCF5 \geq 0$ . The red dashed line marks the zero crossing as the step boundary. Range of effect: ~686 to ~714 (approximately -28 units; ~4%—**a substantial step**). *Nonlinearity:* A near-pure step function with a descending direction—high predicted charges at negative PCF5, low predicted charges at positive PCF5. The step is steep and complete within a narrow PCF5 range near zero. All real data (tick marks) is concentrated at the step. PCF5 (Revenue Expense Contrast) captures the relationship between water utility revenue and expenses. Negative PCF5 identifies utilities where revenue capture is weak relative to expenses—costs exceed revenues or revenue is disconnected from service costs. Positive PCF5 identifies utilities with strong revenue capture exceeding expenses. The step-function result is striking: **utilities with poor Revenue Expense Contrast charge more, not less**. This appears paradoxical, but it is consistent with a cost-recovery rate-setting model under regulatory pressure: utilities that are failing to recover costs through revenue are forced to raise rates to close the gap. The step at zero represents the threshold between cost-recovery shortfall and surplus—utilities on the wrong side of that boundary are charging premium rates as a corrective mechanism.

**PCF7 (Debt Leverage)** A negative step-function: flat high plateau (726) for  $PCF7 < -0.02$ , steep drop between  $PCF7 \approx -0.02$  and 0, flat low plateau (680) for all  $PCF7 \geq 0$ . The red dashed line at zero marks the precise step boundary. Range of effect: ~680 to ~726 (approximately -46 units; ~6.5%—**the largest marginal range of any financial feature in this model**). *Nonlinearity:* An even sharper and higher-amplitude step than PCF5, again with a descending direction. Negative PCF7 (high total debt relative to current revenue) predicts dramatically higher charges than positive PCF7. The magnitude—46 units—makes this the most impactful financial variable in the model. PCF7 (Debt Leverage) is specifically defined by total debt outstanding relative to revenue generation. Negative PCF7 identifies utilities carrying heavy

debt burdens relative to their revenue base—the high-leverage utilities. The finding that high debt leverage predicts higher charges is entirely consistent with public finance theory: debt-laden public utilities must set rates to service their obligations. Municipal bonds and government loans for water infrastructure create binding annual payment obligations that are recovered through rates. The step at PCF7 = 0 represents the threshold between debt-burdened and debt-light utilities. The 46-unit charge premium for debt-burdened utilities is the largest quantified financial effect in any of the four models. Combined with PCF7's H-statistic rank-2 position, this variable is not only a strong direct driver but also a major moderator—debt burden likely amplifies demographic and state effects on charges, meaning the highest charges occur where debt leverage intersects with poverty concentration and West Coast geography.

**PCFS8** (Utility Tax Contribution per Capita) A negative sigmoid: flat high plateau (699) for PCFS8 < -0.01, descending sigmoid through the zero crossing, reaching a flat low plateau (681) for PCFS8 > +0.015. The red dashed line at zero marks the inflection point of the descent.

Range of effect: ~681 to ~699 (approximately -18 units; ~2.5%). *Nonlinearity:* A genuine sigmoid (S-curve) descent, unlike the pure step functions of PCF5 and PCF7. The transition has a meaningful width (spanning roughly -0.01 to +0.015) and is smooth rather than vertical. PCFS8 measures how much tax and general revenue the utility contributes or generates per resident. Negative PCFS8 identifies utilities where per-capita tax and revenue contributions are relatively low (the utility draws less from general tax capacity), while positive PCFS8 identifies utilities contributing more to general fund revenue per capita. The negative direction—low per-capita tax contribution associated with higher charges—may seem counterintuitive, but it reflects the structure of public utility finance: utilities that are not heavily integrated into the general tax base rely more on user charges to fund operations. Utilities deeply embedded in municipal tax structures (positive PCFS8) can draw on broader fiscal resources, allowing lower direct user charges. The smooth sigmoid shape, contrasting with the step functions of PCF5 and PCF7, suggests the tax contribution effect is more graduated and proportional—reflecting that general fund integration is not an all-or-nothing characteristic but a continuum.

**PCD1 × PCD3** (Poverty Gradient × Health Insurance Gap) The joint PDP reveals a **superadditive interaction**: the highest predicted charges occur when both PCD1 (poverty) and PCD3 (health insurance gap) are simultaneously elevated, and the combination produces charges substantially higher than either dimension alone would predict.

- The gradient runs diagonally from lower-left (low both: ~676) to upper-right (high both: ~754), with the contour lines angling roughly 45 degrees—indicating that both dimensions contribute approximately equally to the joint prediction.
- At high PCD1 and high PCD3 (upper-right), predicted charges reach 754, which substantially exceeds the maximum predicted by either variable's 1D PDP alone (729 for PCD1, ~716 for PCD3). This superadditivity—approximately +25 units above what additive effects would suggest—confirms a genuine multiplicative interaction.
- At low PCD1 and high PCD3 (upper-left), charges are moderate (676–692); at high PCD1 and low PCD3 (lower-right), charges are also moderate (707–723). Neither high-poverty-alone nor high-uninsured-alone reaches the extreme values of the joint high-high quadrant.
- The contour structure shows that PCD1 is the primary organizing axis (contours are more compressed horizontally than vertically), consistent with its higher marginal importance and H-statistic.

**Interpretation:** Poverty and health vulnerability in public utility service areas are not merely additive risk factors for high charges—they compound each other. The most financially burdened public utility ratepayers are those living in communities that are simultaneously poor *and* medically vulnerable. This compound disadvantage is likely explained by the infrastructure economics: aging systems in persistently impoverished communities require high per-capita capital investment, while absent or under-resourced health infrastructure makes these communities less able to advocate for or afford rate relief. The superadditive interaction is a policy-critical

finding: addressing only poverty or only health vulnerability in isolation will not fully resolve the charge burden on the most disadvantaged communities.

**PCD1 × PCD4** (Poverty Gradient × Renting Poor vs. Owning Poor) PCD1 is the dominant axis, but PCD4 provides meaningful amplification at high PCD1 values—owning-poor communities served by high-poverty public utilities face a higher charge premium than renting-poor communities with equivalent poverty levels.

- Contour lines are primarily vertical (organized by PCD1), confirming PCD1 dominance, but the contours lean slightly toward the upper-right, indicating PCD4 amplification at high poverty levels.
- At low PCD1 (left side), PCD4 has essentially no effect on charges—the contours are horizontal across the PCD4 axis in the left portion of the plot. The poverty gradient must first be elevated before housing tenure structure matters.
- At high PCD1 (right side,  $PCD1 > +2$ ), the charge premium associated with positive PCD4 (owning poor) becomes visible: predicted charges reach ~742 at high PCD1 + high PCD4 compared to ~716 at high PCD1 + low PCD4.
- The interaction is weaker than  $PCD1 \times PCD3$  but directionally meaningful.

**Interpretation:** Public utility charges are highest in communities that combine extreme poverty with owner-occupied (non-renter) housing—the "rural poor homeowner" archetype. This makes infrastructure sense: rural homeowner communities tend to have dispersed service networks with high per-connection costs, aging private service lines requiring owner-side maintenance, and limited economies of scale in treatment and distribution. The renter-poor communities (low PCD4) at equivalent poverty levels face lower charges, likely because they are served by denser urban or suburban systems with better cost efficiency. The interaction activating only at high PCD1 values confirms that housing tenure structure matters for public utility pricing only in the context of genuine poverty concentration.

**PCF3 × PCFS7** (Taxation Profile × Asset Utilization per capita) This 2D PDP reveals a **quadrant structure** with four near-distinct charge levels, creating the clearest four-way interaction regime observed across all models in this series.

- **Quadrant I (PCF3 < 0, PCFS7 > 0, upper-left):** Teal/cyan color, predicted charges ~695–704. Utilities with low/normal taxation profiles but high per-capita asset utilization (low overhead and revenue efficiency per person).
- **Quadrant II (PCF3 > 0, PCFS7 > 0, upper-right):** Bright yellow, predicted charges ~704–716+. The highest-charge quadrant: utilities with elevated taxation profiles AND high per-capita asset utilization. Tax-prominent utilities that are also efficiently utilizing assets per capita charge the most.
- **Quadrant III (PCF3 < 0, PCFS7 < 0, lower-left):** Deep purple, predicted charges ~668–684. The lowest-charge quadrant: utilities with low taxation profiles and low per-capita asset utilization. These are likely small, financially modest utilities with minimal tax exposure and limited per-capita infrastructure activity.
- **Quadrant IV (PCF3 > 0, PCFS7 < 0, lower-right):** Green/teal, predicted charges ~688–696. Intermediate: elevated taxation but low per-capita asset utilization. Tax-prominent but asset-light utilities.
- The PCF3 boundary (near zero) creates the sharpest division in the plot—essentially a vertical wall between the left two quadrants and the right two quadrants. The PCFS7 boundary (near zero) creates a secondary horizontal division, but the PCF3 boundary is more decisive.

**Interpretation:** The taxation profile of a public utility (PCF3) is the primary organizer of this interaction, with asset utilization per capita (PCFS7) providing secondary modulation. Utilities where public utility taxes are an outlier—either very high or very low—charge more than operationally typical utilities, particularly when combined with high per-capita asset utilization. The upper-right quadrant (high tax + high per-capita utilization) represents large, tax-prominent public utilities that are actively utilizing their infrastructure base per resident: likely metropolitan utilities in high-cost states. The lower-left quadrant (low tax + low utilization)

represents small, fiscally conservative utilities that avoid both tax prominence and heavy infrastructure spending per capita—the lowest-charge archetype. This interaction has no parallel in the private utility models, where taxation operated through PCF2 as a binary discriminator rather than through the combined taxation-utilization structure seen here.

### *E2.1.1 Synthesis: Conceptual Model of Public Utility Charge Drivers*

#### *Tier 1: West Coast State Premium (Dominant Geographic Override)*

**Washington, California, and Oregon** together create the largest and most consistent positive charge premiums in the model, with Washington showing SHAP values reaching +50 and California +40. These represent the structural pricing environment of Western public utilities: higher infrastructure investment requirements, water scarcity management costs, environmental compliance, and regional regulatory frameworks. Texas shows a secondary positive effect, heterogeneous across utilities. **Mississippi and Louisiana** create the dominant negative effects (discounts), consistent with all prior models.

#### *Tier 2: Poverty and Vulnerability Gradient (Continuous, Compounding)*

**PCD1 (Economic Affluence)** is the strongest continuous predictor and the dominant interaction variable (H-statistic  $\sim 0.026$ ). Unlike every other model in this series, PCD1 operates continuously without threshold discretization—public utility pricing responds proportionally to poverty concentration across the entire spectrum. **PCD3 (Health Insurance Gap)** and **PCD4 (Renting Poor vs. Owning Poor)** provide secondary demographic effects, with PCD3 showing a meaningful zero-crossing acceleration and PCD4 a gentler threshold. The  $PCD1 \times PCD3$  superadditive interaction is the most important compound effect in the model: communities that are simultaneously poor and medically vulnerable face the highest public utility charges of any demographic configuration.

#### *Tier 3: Debt Leverage and Revenue Recovery (Public Finance Drivers)*

**PCF7 (Debt Leverage)** provides the largest single financial effect (–46 units, step at zero): high-debt public utilities charge substantially more to service their obligations. **PCF5 (Revenue Expense Contrast)** adds a secondary step effect (–28 units): utilities failing to recover costs through revenues compensate through higher charges. These two financial variables capture the public finance logic of cost-recovery rate-setting—obligations and shortfalls drive prices up in ways that are structurally different from investor-owned utility rate structures.

#### *Tier 4: Tax Structure and Asset Utilization (Context-Dependent Financial Effects)*

**PCF3 (Taxation Profile)** and **PCFS7 (Asset Utilization per capita)** create a four-quadrant interaction structure where the joint effect of tax prominence and per-capita infrastructure activity determines charges in ways that neither variable captures alone. **PCFS8 (Tax Contribution per capita)** provides a smoothly sigmoid moderating effect, with lower tax-base integration associated with higher charges.

#### *Tier 5: Urban Mobility and Secondary Demographics (Background Effects)*

**PCD2 (Urban Mobility & Tenure)** has the most nearly linear PDP of any feature across all models, contributing a modest +15 unit range without threshold effects. PCD5, PCD6, and PCD7 appear in the top-20 importance list but are not among the primary structural drivers, suggesting they capture residual demographic heterogeneity beyond the primary poverty and health vulnerability axes.

### *E2.1.2 Conclusions*

1. **The poverty-charge paradox in public utilities demands policy attention.** Public utilities serving the poorest communities charge more, not less, and this relationship is continuous and accelerating. This is not a threshold effect that assistance programs can resolve by targeting a specific poverty level—it accumulates across the entire poverty spectrum. Targeted rate relief mechanisms (lifeline rates, income-based assistance programs) for public utility ratepayers should be calibrated to the continuous poverty gradient rather than using binary eligibility thresholds.
2. **The PCD1 × PCD3 superadditive interaction identifies the highest-priority communities.** Public utilities serving areas that are simultaneously poverty-concentrated and health-insurance-deficient face predicted charges ~25 units above what additive effects would suggest. These communities—often rural, low-income, and medically underserved—should be prioritized in any federal or state utility assistance framework. The interaction effect means that addressing poverty and health vulnerability together produces more than twice the benefit of addressing either alone.
3. **Debt leverage (PCF7) is the largest single financial driver of public utility charges.** The 46-unit charge premium for high-leverage utilities underscores the role of infrastructure debt in rate-setting. Federal and state infrastructure financing programs that provide grants rather than loans to the highest-leverage utilities would reduce this premium most efficiently. The binary step at PCF7 = 0 means there is a specific threshold of debt-to-revenue ratio that triggers the premium—utilities near that threshold represent the highest-return targets for debt relief.
4. **Revenue Expense Contrast shortfalls (PCF5) represent a secondary but distinct charge-inflation mechanism.** Utilities failing to recover costs through revenues are raising rates to compensate. This pattern warrants investigation: are these utilities in a rate-recovery correction cycle (appropriate), or do they reflect chronic revenue shortfalls due to structural underpricing or demand decline? The binary step again suggests a threshold-based policy target is appropriate.
5. **The West Coast state premium is geography-driven and partially beyond local utility control.** Washington, California, and Oregon's strong positive SHAP effects reflect regional pricing environments that are real but partially reflect genuine cost differences (water scarcity, infrastructure complexity, environmental compliance). Rate benchmarking that does not control for West Coast geography will systematically misclassify these utilities as overpriced when they may be appropriately cost-recovering.
6. **The tax contribution–asset utilization interaction (PCF3 × PCFS7) should guide utility archetype classification.** The four-quadrant structure of this interaction provides a practical basis for classifying public utilities into charge regimes: high-tax/high-utilization (premium), high-tax/low-utilization (moderate), low-tax/high-utilization (moderate-low), low-tax/low-utilization (lowest). Regulatory oversight intensity could be calibrated accordingly.

E2.2 10 year Health Violations



Figure E7 Random Forest Model Results for 10 year health violations of Public Water Utilities

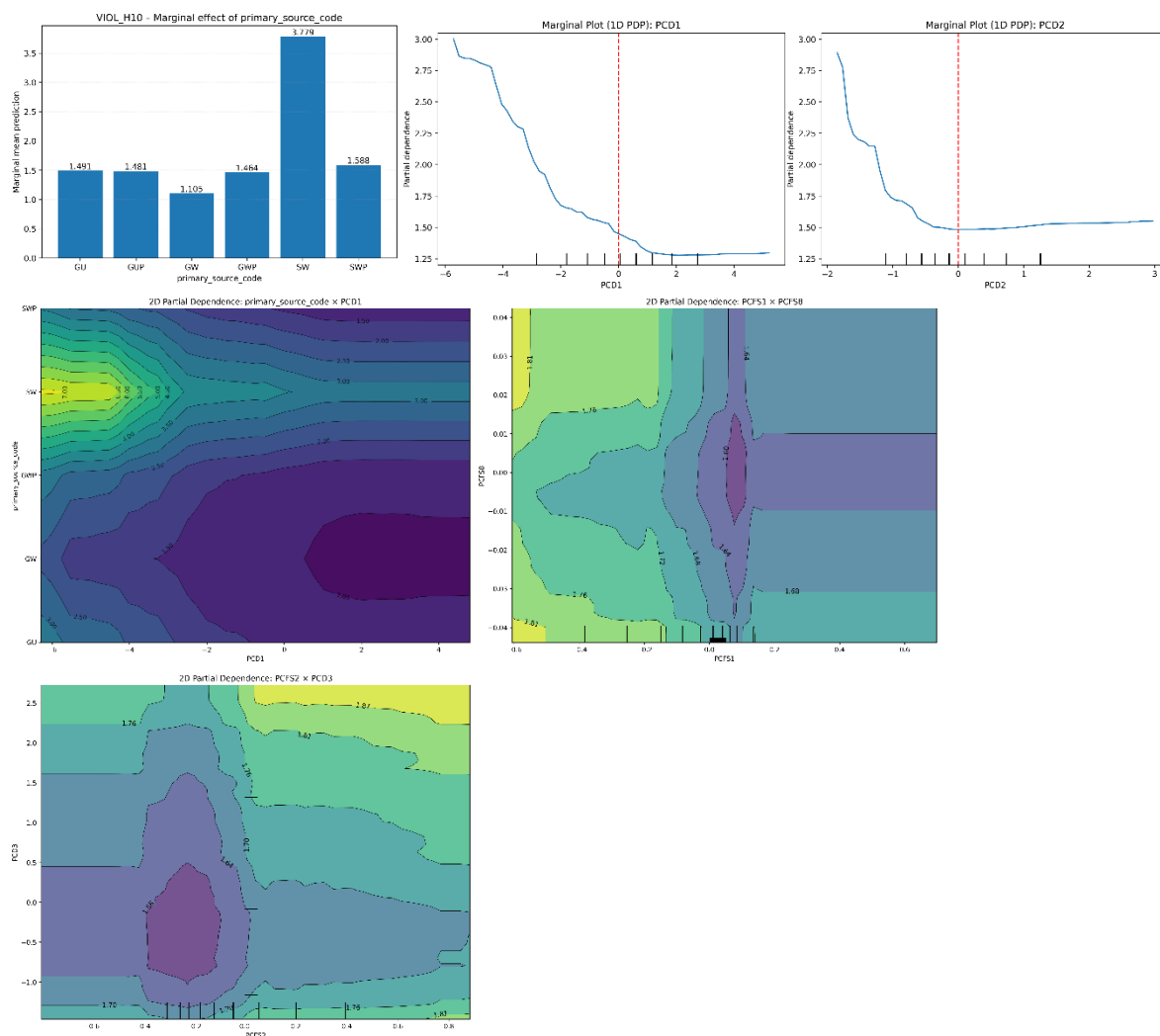


Figure E8 Partial Dependence Plots for Random Forest Model Results for 10 year health violations of Public Water Utilities

A Random Forest (RF) model was trained to predict 10-year health violation counts for **public water utilities** across the United States. The predictor set is the same as the public utility charges model: state indicators, demographic PCs (PCD1–PCD8), aggregate financial PCs (PCF1–PCF10), per-capita scaled financial PCs (PCFS1–PCFS10), population served, and—critically—the primary water source code (groundwater GW, groundwater purchased GWP, groundwater under influence GU, groundwater under influence purchased GUP, surface water SW, and surface water purchased SWP).

This model produces the most structurally concentrated set of findings in the entire series of eight models analyzed. **Water source type is the overwhelming dominant force**, both as a marginal predictor and as the leading interaction variable by an extraordinary margin—its Friedman H-statistic ( $\sim 0.200$ ) is roughly 8 times larger than the next-highest feature (PCD1 at  $\sim 0.025$ ) and is the largest interaction measure observed across any model in this series. **Oklahoma returns as the single most important state** by a wide margin, reprising its dominance from the full-utility violations model but now even more concentrated. **PCD1 (Economic Affluence) reverses direction from the charges model**: where poverty predicted higher charges, here it predicts more violations—but the relationship is now a monotone decreasing function, meaning poorer communities have more violations. The financial structure variables that dominated the private utility violations model are demoted to secondary roles. The overall picture is one where **source**

**type, state identity, and poverty gradient form a tightly interlocked primary triad** that generates the vast majority of violation risk, with financial and demographic secondary features providing conditional modulation.

### Feature Importance Overview

The top-ranked features are:

| Rank | Feature                  | Category              | Interpretation                       |
|------|--------------------------|-----------------------|--------------------------------------|
| 1    | state_OK                 | Geographic            | Oklahoma state indicator             |
| 2    | primary_source_code_GW   | Source type           | Groundwater source                   |
| 3    | PCD1                     | Demographic           | Economic Affluence                   |
| 4    | primary_source_code_SW   | Source type           | Surface water source                 |
| 5    | state_LA                 | Geographic            | Louisiana state indicator            |
| 6    | PCD2                     | Demographic           | Urban Mobility & Tenure              |
| 7    | PCFS2                    | Financial (scaled)    | Debt-Funded Growth per capita        |
| 8    | PCD6                     | Demographic           | Low-Income Edge Cases                |
| 9    | PCD7                     | Demographic           | Income Inequality Gap                |
| 10   | PCD4                     | Demographic           | Renting Poor vs. Owning Poor         |
| 11   | PCFS9                    | Financial (scaled)    | Construction vs. Leverage per capita |
| 12   | PCF1                     | Financial (aggregate) | Overall Utility Scale                |
| 13   | PCFS8                    | Financial (scaled)    | Tax Contribution per capita          |
| 14   | primary_source_code_SW P | Source type           | Surface water purchased              |
| 15   | PCD3                     | Demographic           | Health Insurance Gap                 |
| 16   | PCF4                     | Financial (aggregate) | Investment vs. Maintenance           |
| 17   | PCF2                     | Financial (aggregate) | Revenue/Expense Volume               |
| 18   | PCF7                     | Financial (aggregate) | Debt Leverage                        |
| 19   | PCFS10                   | Financial (scaled)    | Post-Build Recovery per capita       |
| 20   | PCD5                     | Demographic           | Vehicle Dependency                   |

The architecture of this importance ranking has three defining features. First, **state\_OK at rank 1 with MDI  $\approx 0.29$**  is nearly four times the MDI of the second-ranked feature—a more extreme concentration than even Oklahoma's dominance in the full-utility violations model. This signals that Oklahoma's public utility violation environment is qualitatively different from the national baseline, beyond what any combination of source

type, demographics, or financial variables can explain. Second, **source type appears at ranks 2, 4, and 14** (GW, SW, and SWP), confirming water source is a multi-faceted and ubiquitous driver of violations. Third, the **financial variables are present but secondary**—PCFS2, PCF1, PCFS8, PCF4, PCF2, PCF7, and PCFS10 all appear in the top 20, but none approaches the importance level of state\_OK, source type, or PCD1. This mirrors the full-utility violations model structure (source type and state dominating) and contrasts sharply with the private utility violations model (where financial structure dominated).

Two features visible in the SHAP beeswarm but not explicitly in the importance plot warrant note: **population\_served\_count** and **waterutilexp\_rev** (water utility expenditure to revenue ratio) both show meaningful SHAP distributions, confirming that system scale and fiscal balance contribute to violations beyond the ranked features.

The SHAP beeswarm is one of the most directionally clear plots in the entire series:

- **state\_OK**: All red (high-value) dots generate uniformly large positive SHAP values extending from +0.5 to +2.5—the widest and most consistent rightward distribution in the model. Oklahoma public utilities are predicted to have dramatically more violations than their source type, demographics, and financial characteristics would suggest. No Oklahoma utility falls near the zero baseline. The effect is so large it creates a visible gap in the beeswarm between Oklahoma and all other features.
- **primary\_source\_code\_SW (Surface Water)**: High-value (red) dots generate positive SHAP (+0.1 to +0.5), while low values cluster near zero. Surface water public utilities have more violations than non-surface-water utilities—but the effect, while consistent, is modest relative to Oklahoma.
- **PCD1 (Economic Affluence)**: A continuous, mixed distribution centered near zero, with high-value (red, low poverty/high affluence) dots modestly to the left (negative SHAP) and blue dots also appearing. The directionality is subtle in the beeswarm but confirmed by the PDP: higher poverty (more negative PCD1 in this axis definition) = more violations. The SHAP spread is moderate.
- **primary\_source\_code\_GW (Groundwater)**: High-value (blue) dots generate strongly negative SHAP (−0.3 to −0.5)—groundwater public utilities have substantially *fewer* violations than non-groundwater utilities. This is the opposite direction from the full-utility violations model, where groundwater was the highest-risk source. The resolution of this apparent contradiction lies in the 2D interaction with PCD1, discussed below.
- **state\_LA (Louisiana)**: All positive SHAP (+0.1 to +0.5), all red—Louisiana public utilities consistently have more violations than baseline. The effect is uniform but smaller than Oklahoma.
- **state\_TX (Texas)**: Modestly positive SHAP, concentrated within a narrow range (~+0.1 to +0.3). Texas public utilities have slightly elevated violation rates.
- **PCD2 (Urban Mobility & Tenure)**: Negative SHAP for most values, spread leftward—urban mobility (transit-dependent, renter-heavy populations) is associated with fewer violations in public utilities. More urban, rental-dense communities served by public utilities have better compliance records.
- **PCD4 (Renting Poor vs. Owning Poor)**: Positive SHAP for some values—consistent with the charges model finding that owning-poor communities face worse outcomes from public utilities.
- **population\_served\_count**: Modest distribution centered near zero with slight positive SHAP for high values—larger public utility populations served are associated with marginally more violations, consistent with the private utility finding but much weaker here.
- **waterutilexp\_rev**: Red (high-value) dots generate negative SHAP—public utilities with high water utility expenditure relative to revenue have fewer violations. Financially invested public utilities comply better.
- **PCFS2 (Debt-Funded Growth per capita)**: Near-zero SHAP, suggesting this financial variable contributes to importance through interactions rather than direct marginal effects.

The state SHAP maps are striking in their geographic concentration. The direction map shows **Oklahoma as near-black with mean SHAP  $\approx +1.2$ – $1.5$** , accompanied by Louisiana (moderate red,  $\sim +0.3$ – $0.5$ ) and Texas (lighter red,  $\sim +0.1$ – $0.2$ ). New Mexico shows a faint positive effect. The magnitude map confirms Oklahoma as the sole black state (highest magnitude by far), with Texas deep red (second magnitude), Louisiana black-to-dark-red, and West Virginia appearing in the upper right at elevated magnitude—an unexpected appearance of a state not prominent in the importance plot, suggesting some specific compliance challenges in West Virginia's public utility sector.

All other states show near-zero magnitude (pale yellow), confirming that geographic effects are extremely concentrated. The violation geography for public utilities is entirely South-Central: Oklahoma, Louisiana, Texas, and neighboring states form a coherent high-violation cluster that is geographically and hydrogeologically connected (shared aquifer systems, agricultural contamination vectors, similar regulatory histories). The absence of California and Washington from the violation map—despite their dominance in the charges map—is a clean confirmation that what drives charges is categorically different from what drives violations.

#### Interaction Strengths (Friedman H-Statistic)

| Rank | Feature             | H-Statistic (approx.)                 |
|------|---------------------|---------------------------------------|
| 1    | primary_source_code | $\sim 0.200$ (extraordinary dominant) |
| 2    | PCD1                | $\sim 0.025$                          |
| 3    | PCD2                | $\sim 0.005$                          |
| 4    | PCFS2               | $\sim 0.004$                          |
| 5    | PCD4                | $\sim 0.003$                          |
| 6    | PCFS1               | $\sim 0.003$                          |
| 7    | PCD3                | $\sim 0.003$                          |
| 8    | PCF1                | $\sim 0.003$                          |
| 9    | PCD6                | $\sim 0.003$                          |
| 10   | PCF2                | $\sim 0.002$                          |

**The primary\_source\_code H-statistic of  $\sim 0.200$  is the single most extreme interaction measure observed across all eight models in this series**—roughly 8 times larger than PCD1's second-place  $\sim 0.025$ , and approximately 40 times larger than the features at ranks 3–10. This means water source type does not merely contribute to violations additively; it fundamentally reorganizes how every other feature—poverty, geography, financial structure, system scale—translates into violation risk. The 2D PDP of source  $\times$  PCD1 will document this interaction directly. PCD1 at rank 2 repeats its pattern from the charges model of being the dominant demographic interaction variable, but here it operates in its inverse direction. PCFS2

(Debt-Funded Growth) at rank 4 is notable: despite its modest marginal SHAP, it interacts meaningfully with source type and demographic context to modulate violation risk.

### Marginal Effect of Water Source

The bar chart of marginal predicted violations by source type provides the clearest single summary of the model's primary finding:

| Source Code | Description                            | Marginal Mean Prediction |
|-------------|----------------------------------------|--------------------------|
| GW          | Groundwater                            | 1.105                    |
| GUP         | Groundwater Under Influence, Purchased | 1.481                    |
| GU          | Groundwater Under Influence            | 1.491                    |
| GWP         | Groundwater, Purchased                 | 1.464                    |
| SWP         | Surface Water, Purchased               | 1.588                    |
| SW          | Surface Water                          | 3.779                    |

The violation hierarchy is unambiguous: **raw surface water (SW) utilities have predicted violations 3.4 times higher than groundwater (GW) utilities** and more than double any other source category. The SW marginal prediction of 3.779 stands completely apart from the cluster of other source types (1.1–1.6), which form a relatively compact band. Within that cluster, pure groundwater (GW) at 1.105 is the safest, while surface water purchased (SWP) at 1.588 and all groundwater-under-influence variants (GU, GUP at ~1.49) occupy a middle tier.

**Interpretation:** The raw surface water dominance in public utility violations reverses the pattern from the private utility violations model, where groundwater was the primary risk source. The explanation lies in the different contexts: private utilities tapping groundwater in Oklahoma and similar environments face contamination risks from the aquifer itself. Public surface water utilities face violations primarily through the complexity and variability of surface water treatment—monitoring requirements for turbidity, disinfection byproducts, and microbial contaminants are more extensive for surface water, creating more opportunities for compliance failures. Public utilities drawing directly from rivers, lakes, and reservoirs face the most demanding treatment and monitoring obligations of any source category, and failures in any component of this system generate violations. The purchased surface water category (SWP) at 1.588 has substantially lower violations than direct SW, likely because purchased surface water has already undergone primary treatment by the wholesale provider before the public utility receives it.

### Partial Dependence Analysis

**PCD1(Economic Affluence)** Continuously and steeply decreasing from the most negative PCD1 values (approximately 3.0 at PCD1 = −6) to a flat low plateau (approximately 1.28–1.30) for all PCD1 > +1. The steepest descent occurs between PCD1 ≈ −5 and −2, with a further gentle decline through PCD1 ≈ 0 where the curve nears the plateau. The red dashed line at zero marks a point near the bottom of the descent where the curve is already approaching its floor. Range of effect: ~1.28 to ~3.00 (approximately −1.72 violation units; the **largest marginal range of any continuous demographic feature across all eight models in this series** on the violation scale). *Nonlinearity:* Strongly nonlinear with a concave-down

(decelerating) form—the steepest violation reduction occurs at extreme negative PCD1 values (highest poverty), with diminishing returns as poverty decreases. Above PCD1 = +1, the curve is completely flat: once economic conditions are at or above average, additional affluence provides no further violation protection. The plateau at the positive end and the steep descent at the extreme negative end create an asymmetric shape with the most information content concentrated in the poverty tail. Public utilities serving the most poverty-concentrated communities have dramatically more violations—nearly 2.4 times more than utilities serving average or above-average economic areas. This is a large and continuous protective effect of prosperity that operates entirely in the negative PCD1 domain: the violation premium for poverty is captured between PCD1  $\approx$  -6 and 0, with the highest-poverty areas (most negative PCD1) driving the largest absolute violation counts. The plateau above zero means that middle-class and affluent service areas are indistinguishable in their compliance record—once a community exceeds the poverty threshold, additional economic development provides no compliance dividend. Combined with the charges model finding (where PCD1 had a positive, continuous charge effect), this confirms the compound disadvantage structure: the poorest public utility service areas both pay more and receive worse compliance. This is the same PCD1 axis used in the charges model, where positive values indicated low poverty/high affluence and negative values indicated high poverty. The violation PDP shows decreasing violations as PCD1 *increases* (becomes more affluent)—exactly the opposite direction from the charges effect (where increases in PCD1 predicted higher charges). Public utilities serving more affluent communities charge more AND have fewer violations; those serving poorer communities charge less AND have more violations. This compound disadvantage is the single most important cross-model finding in the public utility analysis.

**PCD2 (Urban Mobility & Tenure)** Steeply decreasing from approximately 2.88 at PCD2 = -2 (most vehicle-dependent, low-rental areas), rapid descent through PCD2  $\approx$  -1.5 to -0.5, then leveling to a flat plateau at approximately 1.50 for all PCD2  $\geq$  0. The red dashed line at zero marks the precise transition to the flat plateau. Range of effect: ~1.49 to ~2.88 (approximately -1.39 violation units—a large marginal range, second only to PCD1 among demographic features). *Nonlinearity:* A pronounced threshold pattern with a steep descent entirely in the negative domain and near-complete saturation above zero. The curve is essentially a left-side ramp that fully converges to a stable floor at the zero crossing. Unlike PCD1 (which has a gradual, broad descent), PCD2's active region is confined to the negative support range—below-average urban mobility is where violations accumulate, and average or above-average transit/rental character eliminates virtually all PCD2 violation risk. PCD2 captures the contrast between transit-dependent urban renters (positive values) and vehicle-dependent, low-rental-rate communities (negative values). Public utilities serving highly vehicle-dependent, low-rental-rate communities—typically dispersed rural or exurban areas—have dramatically more violations than those serving transit-oriented urban communities. This is consistent with the infrastructure scale findings from other models: car-dependent, low-density service areas tend to have smaller, more dispersed public utility systems with less monitoring infrastructure and lower institutional compliance capacity. The complete saturation above zero confirms that once a community crosses into even average transit/renter character, the compliance-risk contribution of PCD2 vanishes—the violation risk is entirely concentrated in the most dispersed, car-dependent end of the spectrum.

**primary\_source\_code  $\times$  PCD1** This interaction plot is the most structurally complex and informationally rich 2D PDP in the entire model series. It reveals that **water source type and poverty gradient do not merely add—they multiply in specific quadrants and cancel in others.**

- **SW (Surface Water) at negative PCD1 (upper-left region):** The maximum predicted violations in the entire model (~7.0) occur here—surface water public utilities serving the most poverty-concentrated communities. The yellow region in the upper-left is extreme, with contours running from 7.0 down

through 6.5, 6.0, 5.5 as PCD1 increases from -6 toward -2. This combination (SW + extreme poverty) is catastrophically violation-prone.

- **SW at positive PCD1 (upper-right region):** Violations decline sharply as PCD1 increases for SW utilities, converging to approximately 3.0–3.5 even at high PCD1 values—still substantially elevated compared to all groundwater source types, confirming SW's inherent compliance challenge regardless of community wealth.
- **SWP (Surface Water Purchased) at negative PCD1:** Follows SW directionally but at lower levels (~1.5–2.5), confirming that purchased surface water protects against the worst SW×poverty outcomes.
- **GW (Groundwater) at negative PCD1 (lower-left region):** Starts elevated (~2.5–3.0) at extreme negative PCD1 but declines rapidly toward 1.0–1.5 as PCD1 increases toward zero. The poverty-groundwater combination is elevated but not catastrophic.
- **GW at positive PCD1 (lower-right purple region):** Violations drop to approximately 1.0—essentially the floor prediction in this model. Affluent communities with groundwater sources have the fewest violations of any combination.
- **GWP, GU, GUP at all PCD1 values:** Intermediate violation levels throughout, following the general PCD1 descent but at higher baseline than pure GW.
- **Convergence at PCD1 ≈ 0:** All source types converge toward a common, lower violation level as PCD1 approaches zero and beyond, suggesting that community economic strength overwhelms source-type differences in compliance risk at and above the poverty mean. This convergence is the clearest evidence that poverty and source type interact multiplicatively rather than additively.

**Interpretation:** The source × PCD1 interaction is the central mechanistic finding of this model. Poverty amplifies the risk of surface water violations far beyond what either factor predicts alone: a surface water public utility serving an extremely poor community has predicted violations approximately 4–5 times higher than a surface water utility serving an affluent community, and approximately 7 times higher than a groundwater utility serving an affluent community. The mechanism is likely layered: poverty-area surface water utilities face more contamination challenges from agricultural and industrial runoff, have less capital for treatment upgrades, employ fewer professional operators, and face less active community oversight. Each of these poverty-related disadvantages compounds the intrinsic monitoring and treatment complexity of surface water systems. The convergence above PCD1 = 0 suggests that economic resources above the poverty threshold are sufficient to manage surface water compliance adequately regardless of the inherent source difficulty.

**PCFS1 × PCFS8 (Revenue Efficiency × Tax Contribution per capita)** The joint PDP reveals a **cross-shaped violation landscape** organized around the zero crossings of both financial variables, creating four quadrants with distinct compliance profiles.

- **Left region (PCFS1 < -0.1), all PCFS8:** Elevated violations (~1.76–1.81+), with the highest values in the upper-left corner (negative PCFS1, positive PCFS8 reaching ~1.81). Low per-capita revenue efficiency combined with high tax contribution per capita is the highest-risk financial configuration in this interaction.
- **Center-right region (PCFS1 ≈ 0), all PCFS8:** A narrow vertical minimum zone with very low violations (~1.58–1.60), particularly in the center of the PCFS8 range. This represents utilities at the revenue efficiency breakeven point.
- **Right region (PCFS1 > +0.1), all PCFS8:** Moderate violations (1.64–1.68) that are higher than the center minimum but lower than the far left. A localized purple minimum pocket at the center of the right zone (~1.60) suggests a specific financially balanced archetype.
- **Bottom-left corner (PCFS1 ≈ -0.6, PCFS8 ≈ -0.04):** The highest violation zone in this interaction (~1.81+, yellow), identifying utilities with both very low per-capita revenue efficiency AND very low per-

capita tax contribution. These are financially marginal utilities that generate neither adequate revenue nor meaningful tax contributions per resident—the most resource-starved archetype.

- The horizontal PCFS8 gradient is modest compared to the vertical PCFS1 gradient: PCFS1 is the primary organizer and PCFS8 provides secondary modulation.

**Interpretation:** PCFS1 (Revenue Efficiency per capita) is the organizing axis: utilities with below-average revenue relative to lean expenditures and tax burdens (negative PCFS1) have consistently more violations than those with strong revenue. This is consistent with the general model finding that financially strained public utilities violate more—but the interaction with PCFS8 adds nuance. *The worst violations occur when low revenue efficiency combines with either high or very low tax contribution: both extremes of the tax contribution dimension amplify the revenue-efficiency risk. The narrow minimum at PCFS1  $\approx 0$  represents a specific balanced state—utilities right at the revenue-efficiency breakeven—that achieves the best compliance regardless of tax contribution level.* This non-monotonic response to PCFS1 (worse at both extremes, best at center) is an important finding: it suggests that the extreme revenue-efficiency configurations in either direction create compliance challenges, not just the deficit side.

**PCFS2 × PCD3 (Debt-Funded Growth × Health Insurance Gap)** This interaction produces a **minimum-in-center structure** where the lowest violations occur at a specific combination of moderate negative PCFS2 and low-to-moderate PCD3, with violations rising in all directions from that minimum.

- **Central minimum (PCFS2  $\approx -0.15$  to  $-0.25$ , PCD3  $\approx -0.5$  to  $+0.5$ ):** The lowest predicted violations in this interaction ( $\sim 1.58$ , deep purple) concentrate in a roughly oval region. These are utilities with moderate per-capita debt-funded growth (some active construction but not at extreme debt levels) serving communities with average health insurance coverage. This "Goldilocks" zone achieves the best compliance.
- **Left region (PCFS2  $< -0.4$ , all PCD3):** Elevated violations ( $\sim 1.70$ – $1.76$ ) regardless of health insurance gap level. Very high per-capita debt-driven construction activity is associated with more violations, possibly due to construction-phase compliance challenges or because highly indebted systems are also aging and under-maintained.
- **Right region (PCFS2  $> 0$ , all PCD3):** Elevated violations ( $\sim 1.76$ – $1.87$ +) throughout, with the highest values at the upper right (positive PCFS2, high PCD3  $\approx +2.5$ ). Utilities with positive PCFS2—those with high per-capita debt driven by active construction—serving communities with large health insurance gaps have the worst compliance in this financial-demographic interaction.
- **Upper-right corner (high PCFS2 + high PCD3):** Yellow region ( $\sim 1.87$ ), the maximum in this interaction. Active debt-funded infrastructure growth combined with health-vulnerable communities creates the highest violation risk in this financial-demographic combination—possibly because rapidly expanding systems serve newly developed areas with transient populations and inadequate oversight.
- **PCD3 axis effect:** Above PCD3 =  $+1.5$ , violations increase regardless of PCFS2 level, confirming that extreme health insurance gaps elevate violation risk independently. At low PCD3 (below mean), the PCFS2 effect dominates.

**Interpretation:** The debt-funded growth × health insurance gap interaction reveals a specific compliance risk associated with utilities expanding their infrastructure (positive PCFS2) while serving communities with high medical vulnerability. This combination—growth under financial stress in vulnerable communities—likely represents public utilities in rapidly developing but economically marginal areas: new subdivisions, expanding rural communities, or utilities taking on new service territory without adequate revenue base. The

central minimum at moderate negative PCFS2 represents the optimal financial configuration: some construction activity (system investment and renewal) but at debt levels that don't compromise operational resources. The finding that very low debt-funded construction (most negative PCFS2) also elevates violations suggests a U-shaped compliance response to construction intensity—moderate investment is optimal, with both underinvestment and overinvestment creating compliance risk.

### *E2.2.1 Synthesis: Conceptual Model of Public Utility Violation Drivers*

#### *Tier 1: Oklahoma's Singular Violation Environment (State Override)*

**State\_OK** is the single most important predictor by a margin of approximately 4:1 over the next feature, and its SHAP values extend to +2.5—the highest individual SHAP magnitudes in any violations model.

Oklahoma's public utility sector has a violation environment that is fundamentally unexplained by source type, demographic poverty, or financial structure. The likely drivers include Oklahoma's geology (widespread groundwater contamination from oil and gas operations, agricultural runoff, and naturally occurring arsenic and uranium), the density of small, underfunded public systems in rural areas, and a regulatory environment that has historically under-resourced drinking water compliance. Oklahoma's position in the South-Central high-violation cluster (with Louisiana and Texas) suggests these states share geological and regulatory characteristics.

#### *Tier 2: Water Source as the Dominant Interaction Organizer*

**Primary\_source\_code** carries the largest interaction H-statistic in the series (0.200) and the clearest marginal effect: SW utilities average 3.779 violations versus 1.105 for GW. But source type is not a simple additive predictor—its 2D interaction with PCD1 shows that the SW penalty is amplified 4–5-fold by extreme poverty and essentially collapses in affluent communities. Source type functions as a **\*\*risk multiplier\*\*** whose magnitude depends critically on the demographic and economic context. GW in affluent communities achieves the lowest violation floor (1.0); SW in impoverished communities reaches the model ceiling (~7.0).

#### *Tier 3: Poverty as Continuous Protective Factor*

**PCD1 (Economic Affluence)** is the strongest continuous predictor, operating as a monotone decreasing function with its entire effect concentrated in the poverty domain ( $PCD1 < 0$ ). The 1.72-unit violation range from most to least poor service areas exceeds the range of any other continuous feature. Combined with the charges model's positive PCD1 effect, this establishes the compound disadvantage: poverty simultaneously predicts higher charges and more violations in public utility service areas. **PCD2 (Urban Mobility & Tenure)** provides a complementary geographic-demographic protection: transit-oriented, rental-dense urban communities have substantially fewer violations, with all protective effect captured in the below-zero PCD2 range.

#### *Tier 4: Financial Structure as Secondary Compliance Context*

**PCFS1 × PCFS8** and **PCFS2 × PCD3** interactions reveal that financial structure conditions compliance in non-linear ways. Revenue efficiency (PCFS1) organizes the financial-compliance landscape with a non-monotone response: violations are highest at both revenue deficit and surplus extremes, lowest at the breakeven point. Debt-funded growth (PCFS2) shows a U-shaped compliance response, with moderate construction investment optimal and both underinvestment and overinvestment elevating violations. These findings are consistent with the public finance framework: financially balanced utilities (neither revenue-starved nor over-leveraged) achieve the best compliance records.

#### *Tier 5: Remaining Demographic and Financial Modifiers*

**PCD4 (Renting Poor vs. Owning Poor)**, **PCD6 (Low-Income Edge Cases)**, **PCD7 (Income Inequality Gap)**, **PCD3 (Health Insurance Gap)**, and multiple financial scaled PCs all appear in the top 20 but at substantially lower importance levels. They capture residual compliance variance that is not explained by the primary triad (source, state, poverty) but contribute meaningfully through interactions documented in the PCFS2 × PCD3 interaction.

### *E2.2.2 Conclusions*

**1. Oklahoma's public utility sector requires emergency-level regulatory intervention.** The SHAP values extending to +2.5 and the MDI of ~0.29 indicate that Oklahoma public utility violations are 1–2.5 units above model-predicted levels after controlling for all other factors. This residual is not explained by source type, poverty, or financial structure—it is a state-specific regulatory and geological phenomenon. Federal resources targeting Oklahoma public utility compliance infrastructure, monitoring capacity, and enforcement could produce the largest per-dollar violation reduction of any intervention in this analysis.

**2. Surface water public utilities serving impoverished communities are a national compliance crisis.** The SW × PCD1 interaction producing predicted violations of ~7.0 at the worst-case combination represents extreme multi-dimensional risk. Policy interventions that address only one dimension (source treatment requirements OR poverty assistance) will be less than half as effective as interventions that address both simultaneously. Federal Safe Drinking Water Act funding should explicitly prioritize SW utilities in high-poverty service areas as a joint eligibility criterion.

**3. The compound disadvantage structure demands integrated affordability and compliance policy.** Public utilities serving the poorest communities charge more (positive PCD1 effect on charges) AND violate more (negative PCD1 effect on compliance). Standard affordability frameworks that focus only on rates miss the violation dimension. Regulatory frameworks should require comprehensive reporting that links rate levels to compliance records for the same utilities, enabling identification of the worst-compound-disadvantage utilities.

**4. Source type should be the primary stratification variable for compliance risk benchmarking.** The 3.4× violation ratio between SW and GW, and the extreme interaction with poverty, means that any national violation benchmarking system that does not stratify by source type will systematically misclassify risk. High-violation SW utilities in poor areas require different benchmarks than low-violation GW utilities in affluent areas.

**5. Financial balance (not financial surplus) is the optimal compliance configuration.** The PCFS1 interaction showing minimum violations at the revenue-efficiency breakeven point suggests that over-extraction of surplus from public utilities (positive PCFS1) can be as harmful to compliance as under-resourced deficits. Rate-setting frameworks should target cost recovery with reinvestment, not surplus extraction.

**6. The PCFS2 × PCD3 interaction identifies rapidly expanding utilities in health-vulnerable communities as a priority for proactive oversight.** High debt-funded growth combined with high health insurance gaps (upper-right corner of the 2D PDP) represents the highest-risk financial-demographic combination. Proactive technical assistance and monitoring during infrastructure expansion phases in these communities could prevent violations before they occur.

### E3. Private

We now move to the data set of private utilities, where the key addition is that the 2022 Financial Reports have been integrated into the analysis.

#### E3.1 Average Water Charges

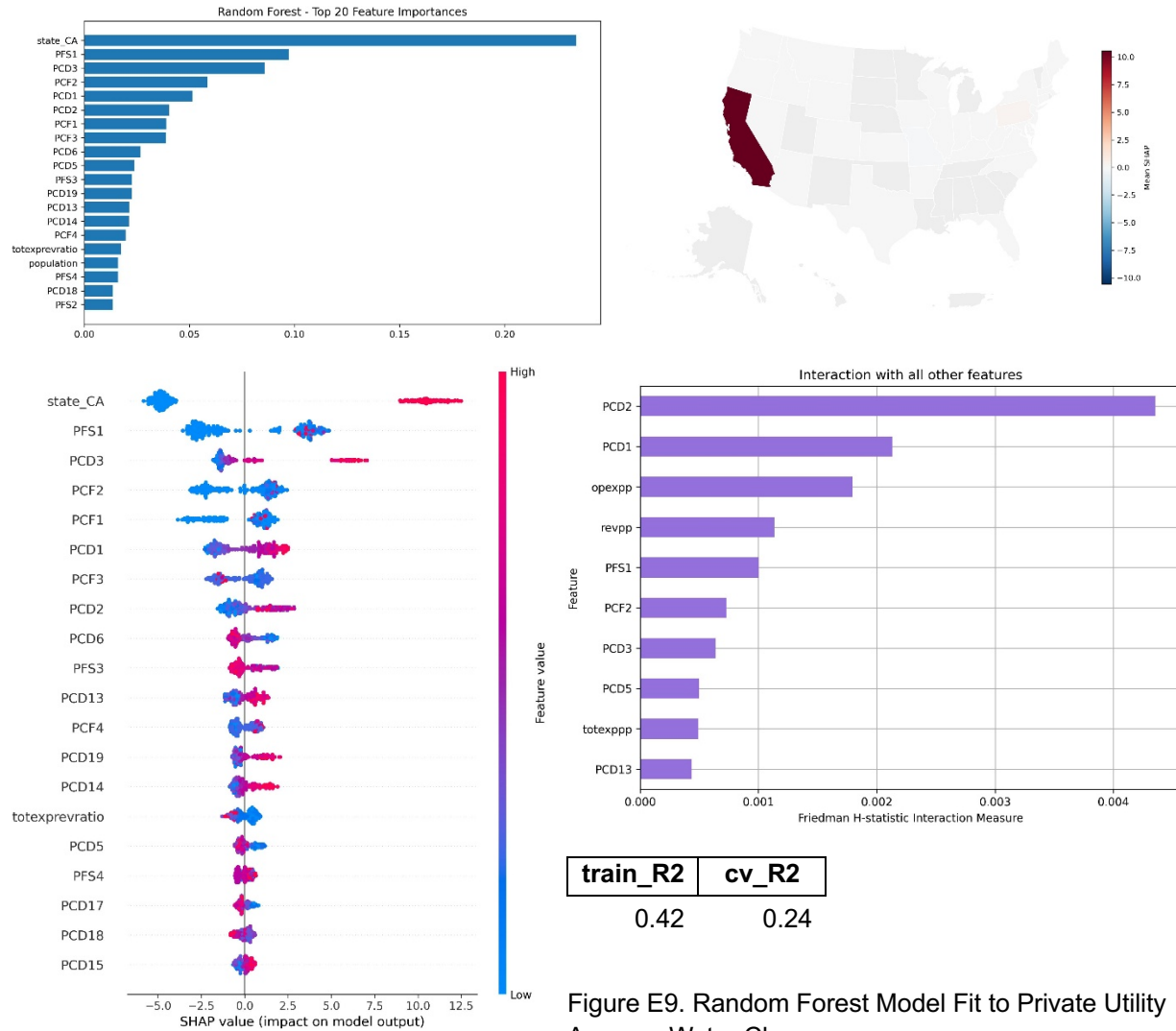


Figure E9. Random Forest Model Fit to Private Utility Average Water Charges

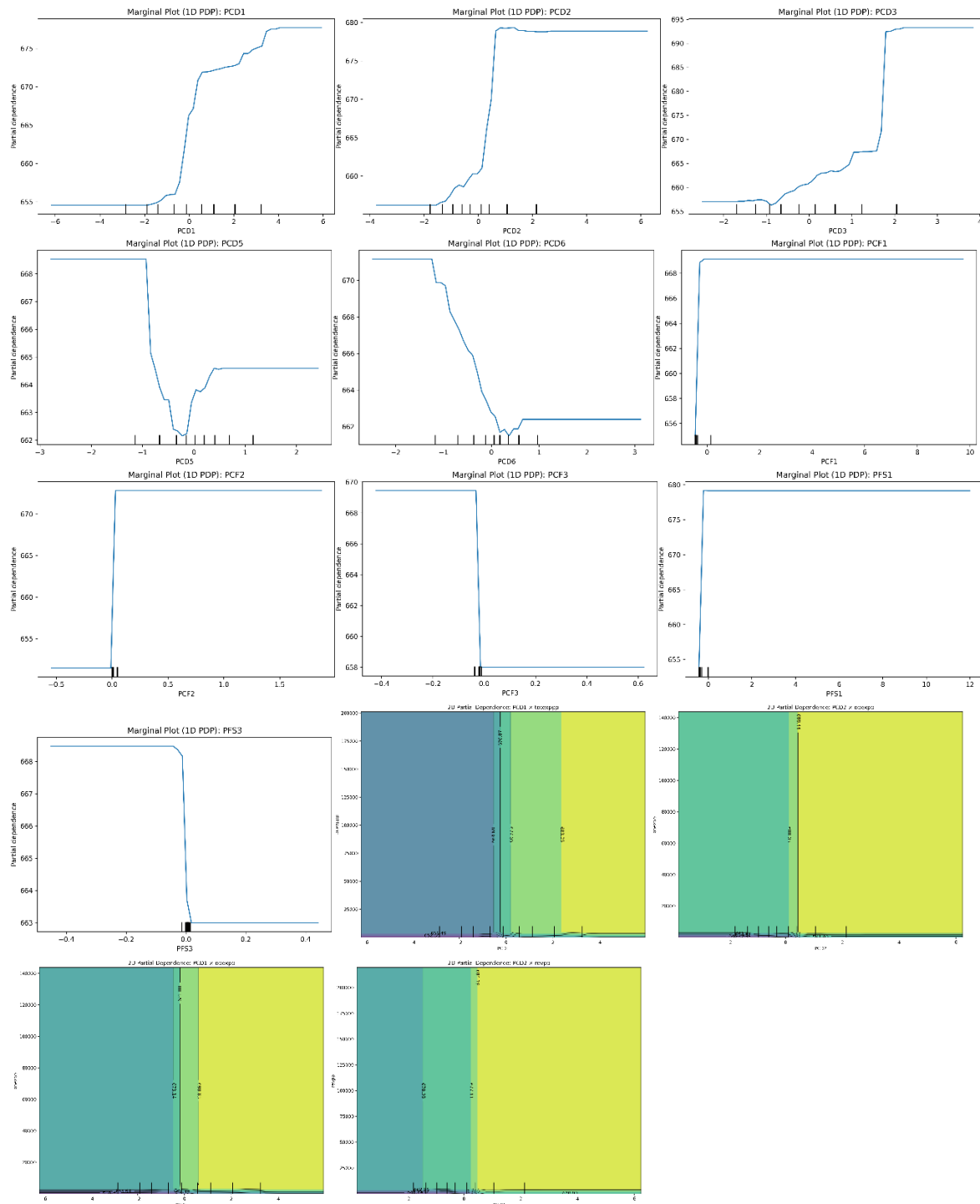


Figure E10. Partial Dependence Plots for Random Forest Model for Private Utilities Average Water Charge

A Random Forest (RF) model was trained to predict average water charges for **private water utilities** across the United States. The model reveals a markedly different charge-setting landscape for private utilities compared to the full utility population. **State identity (California and Pennsylvania) and financial structure dominate**, with the financial PC variables—particularly those capturing system-scale financial volume and per-capita fiscal performance—playing a prominent role that was absent from the EMMA

models. Among demographic predictors, the Knowledge Economy and Urban Diversity dimensions emerge as the most influential, while the deeper poverty and labor-market dimensions that mattered for violations are less prominent here.

### Feature Importance

The top-ranked features are:

| Rank | Feature        | Category              | Interpretation                                    |
|------|----------------|-----------------------|---------------------------------------------------|
| 1    | state_CA       | Geographic            | California indicator                              |
| 2    | PFS1           | Financial (scaled)    | Affluence: avg. per-capita revenue, tax & expense |
| 3    | PCD3           | Demographic           | Knowledge Economy                                 |
| 4    | PCD1           | Demographic           | Economic Affluence                                |
| 5    | PCF2           | Financial (aggregate) | Tax Impact                                        |
| 6    | PCF1           | Financial (aggregate) | Financial Scale                                   |
| 7    | PCD2           | Demographic           | Urban Diversity                                   |
| 8    | PCF3           | Financial (aggregate) | Operational Margin                                |
| 9    | PCD6           | Demographic           | Large Stable Households                           |
| 10   | PFS3           | Financial (scaled)    | Operating Margin per capita                       |
| 11   | PCD13          | Demographic           | Vocational Mid-Income                             |
| 12   | PCF4           | Financial (aggregate) | Expense Gap                                       |
| 13   | PCD19          | Demographic           | Local Wealth Disparity                            |
| 14   | PCD14          | Demographic           | Established Urban Renters                         |
| 15   | totexprevratio | Financial             | Total Expense-to-Revenue Ratio                    |

The most striking departure from both prior models is the **prominent role of financial variables**. PFS1, PCF2, PCF1, PCF3, PFS3, PCF4, and totexprevratio all appear in the top 20—financial features collectively account for a substantial share of total predictive importance. This is consistent with the private utility context: pricing reflects cost structure and revenue strategy in ways that are partially decoupled from demographic context. State\_CA remains the single largest individual feature, consistent with its dominant role in the full-utility charges model.

The SHAP beeswarm plot provides directional interpretation:

- **state\_CA** shows the widest and most right-skewed SHAP distribution: California private utilities command extremely high charge premiums. High-value (red) dots extend well to the right, indicating that California membership is strongly associated with elevated charges regardless of other feature values.

- **PFS1** (Affluence per capita) shows a bimodal SHAP pattern with high values (red) generating moderate positive SHAP and a cluster of blue (low-value) dots near zero. The right-tail positive SHAP for high PFS1 is consistent with higher per-capita financial burden translating into higher charges.
- **PCD3** (Knowledge Economy) exhibits a clear directional SHAP pattern: high values produce positive SHAP (higher charges), while low values cluster near zero. Knowledge-economy service areas pay more—consistent with the full-model finding that affluent, professional communities face higher private utility charges.
- **PCF2** (Tax Impact) shows a notable cluster of high-value (red) dots generating negative SHAP, suggesting that utilities with a tax-dominant financial strategy predict higher charges—potentially reflecting regulatory or rate structures in states where tax burdens are high and rates are constrained.
- **PCF1** (Financial Scale) shows mostly leftward (negative SHAP) concentration for low values, suggesting smaller-volume private utilities charge less on average than high-volume ones.
- **PCD2** (Urban Diversity) generates positive SHAP for high values (right-tail red dots), consistent with urban diversity being associated with higher charges in private utility markets.
- **PCD6** (Large Stable Households) shows scattered positive SHAP for high values—communities with large stable family units in the private utility footprint face modestly elevated charges.
- **state\_PA** (Pennsylvania): while not labeled in the visible top of the beeswarm, the SHAP magnitude map confirms Pennsylvania as the second most geographically distinctive state, consistent with its prominent coloring on the map.

The state SHAP magnitude choropleth is dominated by **two black states: California and Pennsylvania**. Missouri (purple) is the only other state with a meaningfully elevated magnitude. All other states are near-zero (pale yellow to gray). This pattern is more geographically concentrated than the full-utility charges model (which showed broad effects across CA, TX, WA, MS, AK) and far more concentrated than the violations model. The implication is that private utility pricing, once financial and demographic structure is controlled, is largely explained by these state-level regulatory or market environments—California through high cost-of-living and aggressive rate structures, Pennsylvania through its large private utility sector with specific regulatory history.

#### *Interactions (Friedman H-Statistic)*

| Rank | Feature  | H-Statistic (approx.) |
|------|----------|-----------------------|
| 1    | PCD2     | ~0.0043 (dominant)    |
| 2    | PCD1     | ~0.0021               |
| 3    | opexpp   | ~0.0019               |
| 4    | revpp    | ~0.0011               |
| 5    | PFS1     | ~0.0010               |
| 6    | PCF2     | ~0.0008               |
| 7    | PCD3     | ~0.0007               |
| 8    | PCD5     | ~0.0005               |
| 9    | totexppp | ~0.0005               |
| 10   | PCD13    | ~0.0004               |

**PCD2 (Urban Diversity) has the largest interaction H-statistic by a factor of approximately 2**, making it the primary interaction variable despite not being the top marginal predictor. This is the same pattern seen in the EMMA model, where PCD2 topped both the importance and interaction rankings. The financial per-capita variables (opexpp, revpp, totexppp) rank 3rd–5th in interactions—their effect on charges depends strongly on which demographic context the utility operates in. Importantly, the absolute magnitude of all H-statistics is very small (below 0.005), suggesting that while interactions are real, this model is more dominated by main effects and state identity than the previous models were. The financial variables' appearance in the interaction rankings is novel: neither prior model included financial features, and their contextual conditioning by urban diversity and economic affluence is substantively important.

### **PCD1 (Economic Affluence)**

**Shape:** Step-function: flat low plateau (654.7) from  $PCD1 = -7$  through approximately  $-1$ , then a sharp sigmoidal rise between  $PCD1 = -1$  and  $+0.5$ , reaching a second plateau (677–678) that persists through the full positive range. Range of effect:  $\sim 654.7$  to  $\sim 677.7$  (approximately +23 units;  $\sim 3.5\%$  increase).

**Nonlinearity:** A near-pure step function centered near  $PCD1 = -0.5$  to  $0$ . The transition zone is extremely narrow relative to the full support range, and both plateaus are nearly perfectly flat. Economic Affluence is essentially a binary discriminator: utilities either serve economically affluent areas (and charge more) or they do not (and charges are capped at the lower plateau). This could reflect differences in rate-setting philosophy between investor-owned utilities serving wealthy urban markets versus those serving non-affluent areas, or it may reflect state regulatory frameworks that permit higher rates in affluent markets.

**PCD2 (Urban Diversity)** Step-function with flat low plateau (652.3) for  $PCD2 < -1.8$ , a steep ramp between  $PCD2 \approx -1.8$  and  $+0.7$ , and a flat high plateau (678.9) for  $PCD2 > +0.7$ .

Range of effect:  $\sim 652.3$  to  $\sim 678.9$  (approximately +26.6 units;  $\sim 4.1\%$ ). **Nonlinearity:** Very similar to PCD1 in overall form—a two-plateau step function—but with a slightly wider transition zone. The extreme right uptick at very high PCD2 is a novel feature absent from PCD1, possibly reflecting very large, highly diverse urban markets where private utilities operate at premium pricing above even the standard urban-diversity plateau. Urban diversity and its associated wealth concentration creates a persistent charge premium for private utilities. The step function again suggests two operating regimes: non-urban and low-diversity markets form a coherent low-charge cluster, while urban-diverse markets form a high-charge cluster. The separation is cleaner than in the full-utility model, possibly because the private utility population has less geographic heterogeneity within each regime.

**PCD3 (Knowledge Economy)** Step-function: very flat low plateau (657) for  $PCD3 < -1$ , a gradual ramp from  $-1$  to  $+1.7$ , then an extremely steep and nearly vertical jump between  $PCD3 \approx +1.7$  and  $+2.1$ , reaching a high plateau (693) that persists thereafter. Range of effect:  $\sim 657$  to  $\sim 693$  (approximately +36 units;  $\sim 5.5\%$ —the **largest marginal range of any demographic feature in this model**). **Nonlinearity:** A two-stage response: a shallow, modest ramp through most of the support range followed by an extremely sharp, high-amplitude step at the upper tail. The step at  $PCD3 \approx +1.8$ – $2.0$  is the most extreme threshold observed in any of the three models analyzed. The tick marks confirm data density is moderate across the range, so the step is not simply a data-sparsity artifact—it reflects a genuine, concentrated pricing discontinuity at very high knowledge-economy intensity. Knowledge economy concentration—highly educated, professional, technology-sector communities—creates a premium charge structure that activates sharply only at the highest levels of this dimension. The modest slope through moderate knowledge-economy values suggests this factor provides minimal pricing premium in most service areas, but at the extreme high end (likely Silicon Valley, Research Triangle, and similar tech-cluster areas within the private utility service territory), a qualitatively different pricing tier emerges.

**PCD5** (Young Starter Families) Step-function with high plateau (669) for  $PCD5 < -1$ , a sharp downward drop between  $PCD5 \approx -1$  and  $-0.25$ , a narrow trough near  $PCD5 \approx -0.3$  (662.2), then a recovery to an intermediate flat plateau ( $\sim 664.5$ ) for  $PCD5 > 0$ .

Range of effect:  $\sim 662.2$  to  $\sim 669$  (approximately  $-6.8$  units; relatively small). Communities dominated by established, older, or non-family households (low PCD5) face higher private utility charges than communities with young families. This is consistent with the demographic sorting of private utility service areas: older, more established residential markets often overlap with the affluent suburban and knowledge-economy areas that drive the PCD1 and PCD3 effects. Young-family communities served by private utilities tend to be newer developments with more competitive rate structures.

**PCD6** (Large Stable Households) Step-function with flat high plateau (671) for  $PCD6 < -1$ , a steep drop between  $PCD6 \approx -1.2$  and  $-0.2$ , then a flat low plateau (662.2) persisting from  $PCD6 \approx 0$  onward to  $PCD6 > 3$ . Range of effect:  $\sim 662.2$  to  $\sim 671$  (approximately  $-8.8$  units).

Large stable households—the suburban family homeowner archetype—pay relatively less in private utility charges, while communities without this demographic signature (likely more urban or transient) pay more. This is consistent with the common observation that investor-owned utilities serving established suburban customers face stronger rate-payer advocacy and regulatory scrutiny, producing lower margins and thus lower charges, compared to utilities serving higher-turnover urban markets.

**PCF1** (Financial Scale) Nearly pure step-function: very low value for PCF1 at its minimum (below  $-0.3$  on the original scale), then an essentially vertical jump to a flat plateau ( $\sim 669$ ) at  $PCF1 \approx -0.1$ , which persists across the entire range 0 to 10. Only a handful of tick marks appear at the left-boundary transition. Range of effect:  $\sim 655$  to  $\sim 669.2$  (approximately  $+14.2$  units). PCF1 represents the aggregate financial scale of the utility (the average of revenues, operating expenses, and total expenses). The step function implies a minimum operational scale is required to sustain higher charge levels, but once that threshold is exceeded, absolute financial volume does not drive charges up further. Large multi-system private utilities and small single-system ones both operate in similar charge territory once each crosses the viability threshold. This is consistent with regulatory rate-of-return models where allowed returns are tied to rate base rather than absolute revenue scale.

**PCF2** (Tax Impact) Step-function: flat low plateau (652.3) for  $PCF2 < -0.01$ , essentially vertical jump to flat high plateau (672.8) at  $PCF2 \approx 0$ , persisting across all positive values. Range of effect:  $\sim 652.3$  to  $\sim 672.8$  (approximately  $+20.5$  units).

*Nonlinearity:* The sharpest and cleanest step-function in the model—the entire transition occurs essentially at the  $PCF2 = 0$  boundary. The tick marks confirm data density is concentrated right at this boundary. This feature has a binary character: utilities with negative PCF2 (opex-dominant, where operating expenses exceed tax burden) charge substantially less than utilities with positive PCF2 (tax-dominant). PCF2 captures the contrast between tax burden and sales/operating expenses in the financial mix. Utilities where the tax component is relatively dominant charge approximately \$20 more per year on average than utilities where operating costs dominate. This likely reflects that tax-dominant utilities (often in high-tax jurisdictions, possibly Pennsylvania or other regulated states) face a different cost-recovery structure where taxes are explicitly passed through in rates, while opex-dominant utilities (possibly smaller, more cost-efficient, or in lower-tax environments) operate at lower rate levels. The zero-crossing threshold is likely a real structural break in the data rather than a smooth continuum.

**PCF3** (Operational Margin) Step-function: flat high plateau (669.7) for  $PCF3 < -0.08$ , essentially vertical drop at  $PCF3 \approx -0.03$ , flat low plateau (658) for all  $PCF3 > 0$ .

Range of effect:  $\sim 658$  to  $\sim 669.7$  (approximately  $-11.7$  units). Private utilities operating with negative operational margins—spending more on operations than they receive in revenues—charge higher rates. This may reflect utilities that are under-resourced or in the process of rate-increase proceedings, where

high operating costs drive rate petitions. Alternatively, it could reflect that utilities with strong regulatory oversight are more constrained from raising rates even when profitable, while those with cost challenges are granted higher rates. The tick marks confirm that essentially all real data observations sit right at the step, so the PDP behavior outside the observed range should be interpreted cautiously.

**PFS1 (Financial Affluence (Per-Capita Scaled))** Near-perfect step-function: low flat plateau (653.7) for PFS1 at minimum values, steep rise just past the left boundary ( $PFS1 \approx -0.5$  to 0), then flat high plateau (679) that persists from  $PFS1 \approx 0$  across the entire support range to  $PFS1 = 12$ . Range of effect: ~653.7 to ~679 (approximately +25.3 units). *Nonlinearity*: Structurally identical to PCF1—a pure threshold with no gradation above the threshold. The per-capita financial affluence measure (average of per-capita revenue, tax, and expenses) functions as a binary discriminator in this model: utilities with any detectable per-capita financial activity above the minimum floor predict a \$25 higher average charge than those below. PFS1 is the per-capita version of PCF1, and its step-function PDP suggests that both aggregate and per-capita financial scale trigger the same binary pricing regime. This is consistent with the idea that private utility charge levels are determined primarily by which market segment a utility serves (premium vs. non-premium), with financial scale simply being a correlated marker of that segment membership rather than a continuous driver.

**PFS3 (Operating Margin per Capita)** Flat high plateau (669.3) for  $PFS3 < -0.08$  (approximately), then steep drop to flat low plateau (663) at  $PFS3 \approx 0$ , persisting through all positive values. Range of effect: ~663 to ~669.3 (approximately -6.3 units). Consistent with PCF3, negative per-capita operating margin is associated with higher charges. This reinforces the interpretation that cost-pressured utilities—whether measured at the aggregate or per-capita level—operate in higher-rate environments. The effect is smaller here (-6.3 vs. -11.7 units for PCF3), suggesting the aggregate operational margin carries more predictive signal than the per-capita version.

#### **PCD1 × opexpp (Operating Expense per Customer)**

**Key finding:** The 2D PDP reveals that PCD1 (Economic Affluence) and per-capita operating interact only such that very low opexpp leads to a completely different water charge dynamic than above the small threshold where there is a nearly linear increase in water charges with PCD1.

#### **PCD1 × totexppp (Total Expense per Customer)**

**Key finding:** Structurally identical to the PCD1 × opexpp result: near-perfectly vertical contours across the full totexppp range (0 to 200,000), with a stepped boundary near  $PCD1 \approx 0$ .

#### **PCD2 × opexpp (Operating Expense per Customer)**

**Key finding:** Consistent with the previous 2

#### **PCD2 × revpp (Revenue per Customer)**

**Key finding:** Very similar to the previous 3. Revenue per customer modestly amplifies the PCD2 charge effect at the low end: non-urban, low-revenue private utilities are doubly discounted. At high revenue levels, the demographic boundary becomes less sharp—high-revenue utilities can charge near the urban premium even in non-urban markets. This is the clearest evidence in the model that financial performance and demographic context interact, but the interaction is relatively modest and in the range of low income utilities.

### *E3.1.1 Synthesis: Conceptual Model of Private Utility Charge Drivers*

#### *Tier 1: State Regulatory Environment (Dominant Override)*

**California and Pennsylvania** have charge premiums that are demographically and financially irreducible. These represent the largest predictive signals in the model and likely encode the regulatory rate-of-return

frameworks, cost-of-living adjustments, and market structures specific to the investor-owned utility sectors of these states. Pennsylvania's prominence (absent from the full-utility model) is consistent with its large, regulated private water utility sector.

#### *Tier 2: Demographic Market Character (Binary Regime Structure)*

**PCD3 (Knowledge Economy), PCD2 (Urban Diversity), and PCD1 (Economic Affluence)** are the three most important continuous predictors, and all three exhibit step-function relationships. Together they define a binary market segmentation: high-charge private utility markets are characterized by knowledge-economy, urban-diverse, economically affluent service areas; low-charge markets lack all three. The Knowledge Economy dimension has the largest marginal range (+36 units) and activates most sharply at the extreme upper tail, suggesting that Silicon Valley-type markets represent a qualitatively distinct pricing tier.

#### *Tier 3: Financial Structure (Threshold Discriminators)*

**PCF2 (Tax Impact), PCF1 (Financial Scale), PFS1 (Per-Capita Affluence), PCF3 (Operational Margin), and PFS3 (Per-Capita Operating Margin)** all show step-function PDPs. The financial variables effectively sort utilities into pricing tiers—tax-dominant utilities charge more, operationally profitable utilities charge less (counterintuitively), and scale thresholds function as binary classifiers. The 2D PDPs confirm these financial effects are largely **independent of demographic context**: opexpp and totexpp do not interact with PCD1 or PCD2.

#### *Tier 4: Household Structure (Modest Modifiers)*

**PCD5 (Young Starter Families) and PCD6 (Large Stable Households)** both show step-function effects of smaller magnitude (~7–9 units). They primarily define within-tier variation once the Tier-1 and Tier-2 segmentation is established. The direction of both effects is consistent with the interpretation that established, older, larger households in suburban private utility markets face more rate pressure (through advocacy and regulatory engagement) than younger, transient, or urban rental markets.

#### *Tier 5: Interaction Structure (PCD2-Dominated)*

**PCD2's dominant Friedman H-statistic** indicates it is the primary moderator of financial variable effects. The interaction with revpp is the most substantive: high-revenue non-urban utilities can partially close the urban-diversity pricing gap, but the demographic divide persists. The near-total absence of opexpp and totexpp interaction effects is a striking finding: cost structures do not moderate the demographic charge premium.

The decoupling of operating and total expenses from charges—confirmed across both PCD1 and PCD2 interaction plots—is perhaps the most policy-significant finding unique to this model. In the private utility context, **charges are not set by costs**: they are set by market position (demographic tier), regulatory environment (state), and financial structure (Tax Impact). This has direct implications for rate case analysis and prudence reviews: demographic and state indicators should be used alongside cost data when evaluating whether private utility rates are justified.

#### *E3.1.2 Conclusions*

1. **California and Pennsylvania create structurally distinct private utility pricing tiers.** Their dominance in the SHAP map confirms that state regulatory frameworks—not just costs or demographics—are the primary determinants of private utility charge levels. Comparative rate analysis must explicitly account for these state effects rather than treating them as unexplained residuals.

2. **Knowledge economy concentration activates a discrete premium tier.** The extreme step in PCD3 at values above +1.8 identifies a specific utility subpopulation (likely serving tech-cluster and university-adjacent markets) with qualitatively different pricing. Regulatory oversight of rate cases in these areas should account for the unusual leverage that knowledge-economy demographics give private utilities.

**3. Financial structure matters through threshold effects, not cost pass-through.** The step-function PDPs of all financial variables, and the near-complete absence of interaction between cost variables and demographic premiums in the 2D PDPs, indicate that private utility charges are not primarily cost-driven in the statistical sense. Tax-dominant financial structures and scale thresholds function as binary charge regime discriminators.

**4. The counterintuitive negative margin–high charge relationship warrants investigation.** PCF3 and PFS3 both show that utilities with negative operational margins charge more. This could reflect cost-pressured utilities winning rate increases, but it could also reflect accounting or data quality issues. Targeted audits of the most charge-intensive, operationally negative-margin private utilities are warranted.

**5. Household structure dimensions (PCD5, PCD6) provide modest but real differentiation.** Within the high-charge demographic tier, large-stable-household and young-family community profiles show lower charges, consistent with stronger rate-payer engagement in these markets. Regulatory attention to transient, renter-heavy urban private utility markets—where household advocacy is weaker—may be warranted.

### E3.2 10-year Health Violations

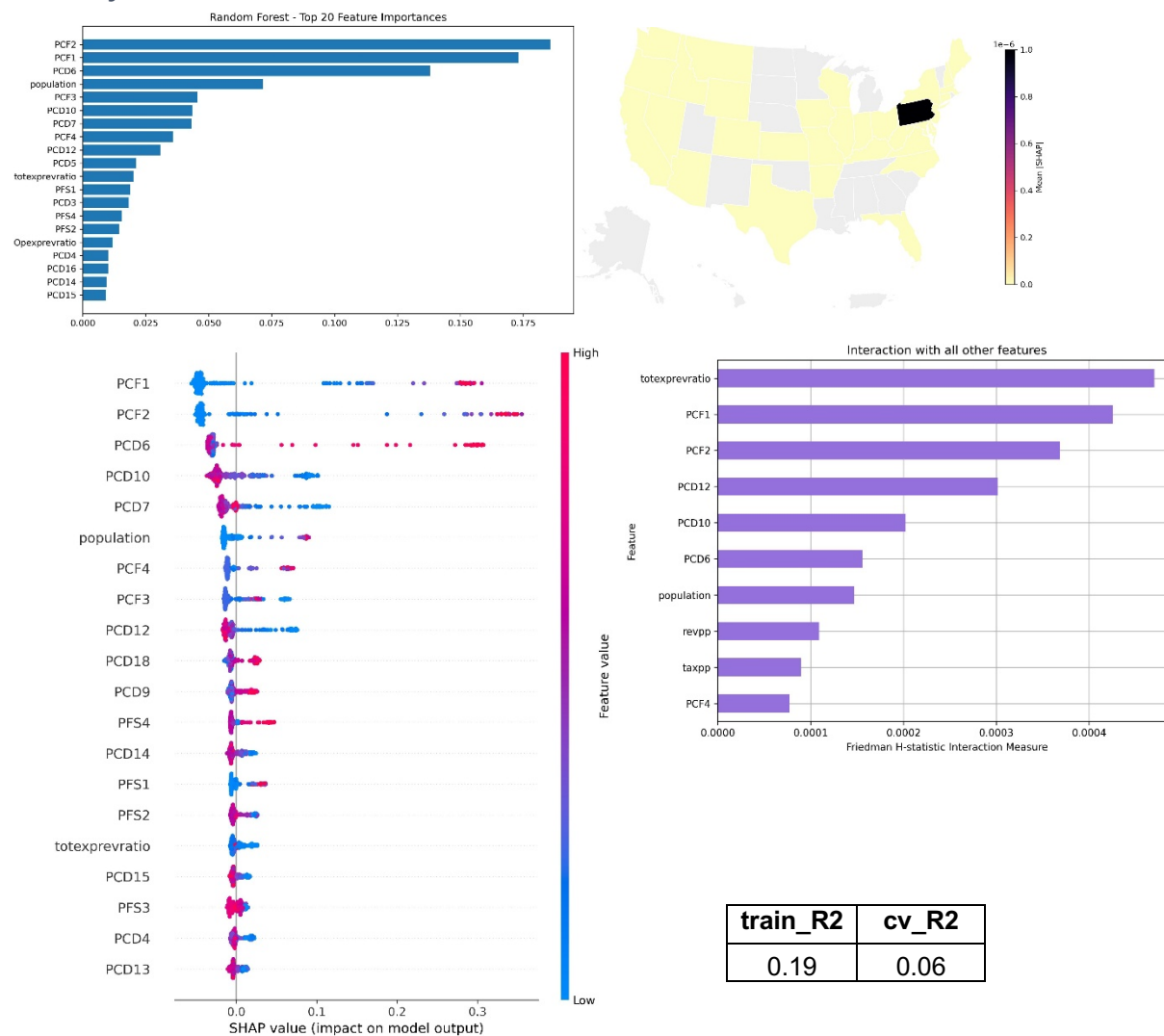


Figure E11. Random Forest Model results for 10-year health violations for private utilities. Note the dramatically lower variance explained.

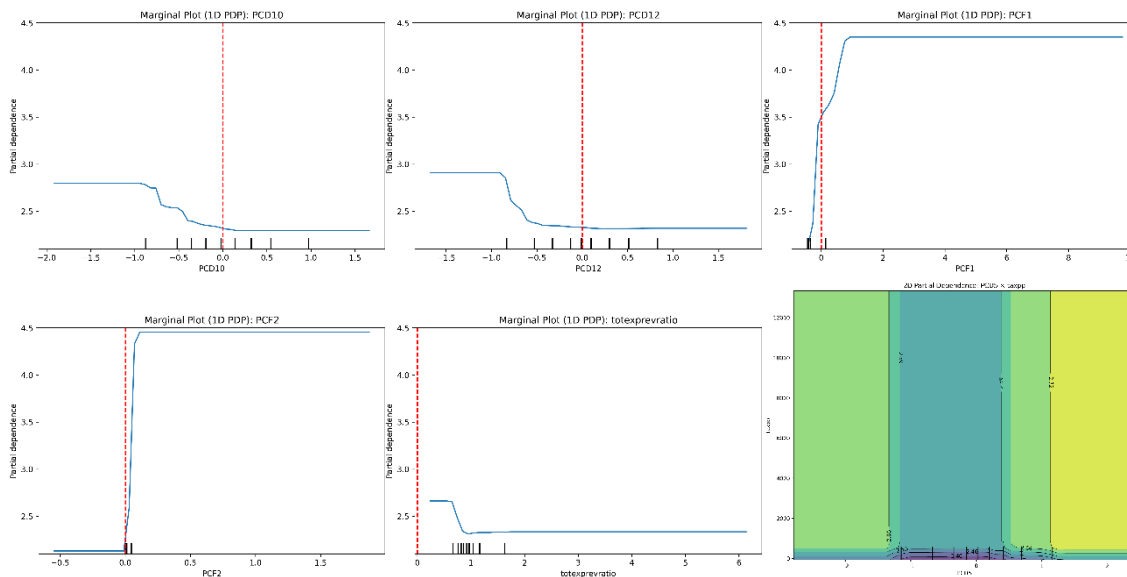


Figure E12. Partial dependence plots for selected variables for 10-year health violations for Private Utilities

A Random Forest (RF) model was trained to predict 10-year health violation counts for **private water utilities** across the United States. This model applies the same private utility sample and financial predictor set used in the private utility charges analysis, enabling direct comparison of what drives pricing versus what drives compliance failure within the same population of investor-owned systems.

The violation model tells a structurally distinct story from private utility charges. **Financial structure dominates overwhelmingly**: PCF2 (Tax Impact) and PCF1 (Financial Scale) are the top two predictors by a wide margin, collectively accounting for the majority of predictive importance—a reversal from the charges model where these were secondary to state identity and demographic character. Note that the aggregation of PWSID Data into companies, combined with the acquisition of large violation public entities by a few of the large private operators led to spurious conclusions on violations with respect to PCF1, and consequently, we dropped those conclusions. This is less of a problem with the EMMA data set where the analysis is at the PWSID level and hence the small, poor attributes at the PWSID level accounted for the correct relationships with scale. Among demographic predictors, **PCD6 (Large Stable Households)** is the third most important variable, a striking elevation from its secondary role in the charges model. State identity is effectively irrelevant for violations in this private utility context: **Pennsylvania is the only state with a detectable SHAP effect**, and even that is modest and narrowly confined. The response variable distributions are on a violation-count scale (roughly 2.3 to 4.5 partial dependence units), and the PDPs—annotated with red dashed lines marking the zero-crossing—reveal that the sign and magnitude of financial PCs relative to zero is the primary organizing principle for violation risk. Nearly all PDPs show sharp step-function or threshold behavior, consistent with the private utility charges model architecture, but the directional logic is largely inverted or orthogonal to what drives charges.

### Feature Importance

The top-ranked features are:

| Rank | Feature | Category              | Interpretation  |
|------|---------|-----------------------|-----------------|
| 1    | PCF2    | Financial (aggregate) | Tax Impact      |
| 2    | PCF1    | Financial (aggregate) | Financial Scale |

| Rank | Feature        | Category              | Interpretation                     |
|------|----------------|-----------------------|------------------------------------|
| 3    | PCD6           | Demographic           | Large Stable Households            |
| 4    | population     | System characteristic | Population served                  |
| 5    | PCF3           | Financial (aggregate) | Operational Margin                 |
| 6    | PCD10          | Demographic           | The Working Poor                   |
| 7    | PCD7           | Demographic           | Active Labor Force                 |
| 8    | PCF4           | Financial (aggregate) | Expense Gap                        |
| 9    | PCD12          | Demographic           | Educational Residuals              |
| 10   | PCD5           | Demographic           | Young Starter Families             |
| 11   | totexprevratio | Financial             | Total Expense-to-Revenue Ratio     |
| 12   | PFS1           | Financial (scaled)    | Per-Capita Affluence               |
| 13   | PCD3           | Demographic           | Knowledge Economy                  |
| 14   | PFS4           | Financial (scaled)    | Financial Balance Measure          |
| 15   | PFS2           | Financial (scaled)    | Tax Ratio                          |
| 16   | Opexprevratio  | Financial             | Operating Expense-to-Revenue Ratio |
| 17   | PCD4           | Demographic           | Educational Attainment             |
| 18   | PCD16          | Demographic           | Educated Financial Stress          |
| 19   | PCD14          | Demographic           | Established Urban Renters          |
| 20   | PCD15          | Demographic           | Immigrant Homeowners               |

The most striking feature of this ranking is the **total dominance of financial variables in the top positions**. PCF2 and PCF1 alone account for an MDI approximately 3–4 times larger than any demographic variable. In the charges model, financial variables were secondary to state identity and key demographic dimensions (PCD3, PCD1, PCD2); here they are primary. The complete absence of state indicators from the top 20 is also remarkable—in the full-utility violations model, state\_OK dominated the rankings. For private utilities, violations are explained by the financial and demographic character of the utility, not by the state it operates in. PCD6 (Large Stable Households) at rank 3 with MDI  $\approx 0.14$  is the highest-ranked demographic feature and notable for its elevation: in the charges model it was a secondary modifier ( $\sim$ rank 9); here it nearly matches PCF1 in importance. Population served appears prominently at rank 4, echoing the scale-protection finding from the full-utility violations model. PCD10 (The Working Poor) and PCD7 (Active Labor Force) are labor-market dimensions that did not appear prominently in the charges

model, suggesting that the socioeconomic character of private utility service areas affects compliance quite differently from how it affects pricing.

The SHAP beeswarm confirms and extends the importance ranking:

- **PCF2 (Tax Impact):** High-value dots generate the furthest rightward SHAP values in the model ( $\sim +0.30$ – $0.35$ ), while low-value dots sit near zero or slightly left. Tax-dominant private utilities are predicted to have substantially more violations than opex-dominant ones. This directional finding contrasts sharply with the charges model, where PCF2 operated as a step-function charge discriminator but in the opposite sense—tax-dominant utilities charged more but also violated more.
- **PCD6 (Large Stable Households):** High-value (red) dots extend substantially rightward to  $+0.3$ , while the distribution is otherwise concentrated near zero. This is a right-skewed positive SHAP pattern: communities with high concentrations of large stable households served by private utilities have markedly elevated violation predictions. This is a reversal from the charges model, where PCD6 was associated with lower charges—large stable households thus appear to pay less but experience more violations from private utilities, suggesting a compliance-affordability trade-off specific to these communities.
- **PCD10 (The Working Poor):** The distribution is almost entirely left of zero—blue (low-value) dots generate negative SHAP. Higher working-poor concentration is associated with fewer violations in private utility service areas. This is counterintuitive: poorer communities typically face more violations (as seen in the full-utility model). The reversal in the private utility context may reflect that private utilities serving working-poor communities operate in simpler, lower-volume service environments where compliance is more tractable.
- **PCD7 (Active Labor Force):** Similar to PCD10—mostly leftward SHAP with blue (low-value) dots extending to  $-0.15$ . Active labor force communities predict fewer violations.
- **population:** Blue (low-value) dots generate positive SHAP, while high-value dots are near zero or slightly negative—confirming that smaller populations served by private utilities are associated with more violations, consistent with the scale hypothesis (even within the private utility population).
- **PCF3 (Operational Margin) and PCF4 (Expense Gap):** Both show modest positive SHAP for high values, suggesting that favorable financial positions (positive margin, larger expense gap) are associated with slightly more violations—a pattern consistent with the charges model finding that financial structure indicators do not protect against violations in the straightforward cost-capacity manner one might expect.
- **PCD12 (Educational Residuals):** Shows a mix of positive (red) SHAP at high values and negative SHAP at low values, but concentrated near zero. Its importance ranking suggests it functions as a contextual modifier rather than a primary driver.
- **totexprevratio:** Mostly concentrated near zero SHAP with a small positive skew for some values—modest and consistent with its position in the middle of the importance ranking.

The state SHAP maps reveal a near-complete absence of state-level effects on violations for private utilities. Only **Pennsylvania** shows a detectable positive mean SHAP (direction map: dark red; magnitude map: black)—all other states are uniform pale yellow or gray, indicating essentially zero residual state contribution. This is a fundamental contrast with both the full-utility violations model (where Oklahoma dominated and Louisiana was secondary) and the private utility charges model (where California and Pennsylvania both contributed strongly). The implication is that private utility violation risk, once financial structure and demographic context are controlled, is **not a state-regulatory phenomenon**—it is a utility-level financial and demographic phenomenon. Pennsylvania's positive violation SHAP echoes its appearance in the charges model (where it also contributed a positive premium), suggesting Pennsylvania's private utility sector is distinctive in both pricing and compliance outcomes, possibly related to the age and complexity of its investor-owned infrastructure.

## Interaction Strengths (Friedman H-Statistic)

| Rank | Feature     | H-Statistic (approx.) |
|------|-------------|-----------------------|
| 1    | totexpratio | ~0.00050 (dominant)   |
| 2    | PCF1        | ~0.00043              |
| 3    | PCF2        | ~0.00037              |
| 4    | PCD12       | ~0.00030              |
| 5    | PCD10       | ~0.00020              |
| 6    | PCD6        | ~0.00015              |
| 7    | population  | ~0.00014              |
| 8    | revpp       | ~0.00010              |
| 9    | taxpp       | ~0.00009              |
| 10   | PCF4        | ~0.00008              |

**totexpratio** leads the interaction ranking, even though it sits only at rank 11 in marginal importance—a pattern diagnostic of a feature whose primary influence is through context-dependence rather than as a main effect. The expense-to-revenue ratio modifies how other features translate into violation risk, meaning the relationship between financial structure and violations differs depending on whether a utility is running above or below a revenue-balanced budget. PCF1 and PCF2 rank 2nd and 3rd in interactions, consistent with their top marginal importance: they both drive violations directly and moderate each other's effects. PCD12 (Educational Residuals) at rank 4 in interactions is a more prominent moderator than its marginal importance (rank 9) suggests, consistent with the hypothesis that educational residuals in the service area change how financial structure translates into compliance behavior. The absolute H-statistic magnitudes are very small (all below 0.001), similar to the charges model, confirming that main effects dominate this private utility model architecture.

## Partial Dependence Analysis

**PCD10** (The Working Poor) Two-plateau step function with a high plateau (2.80) for  $PCD10 < -0.75$ , a gradual decline through the transition zone ( $-0.75$  to  $0$ ), and a flat low plateau (2.30) for all  $PCD10 > 0$ . The red dashed line at  $PCD10 = 0$  marks the precise zero crossing where the descent is nearly complete. Range of effect: ~2.30 to ~2.80 (approximately  $-0.50$  violation units from most to least working-poor areas). Private utility service areas with high concentrations of working-poor households have significantly more violations (~0.5 additional violations over 10 years) than those with working-poor below the mean. This direction is consistent with expectation from the broader literature. Combined with the SHAP finding that PCD10 generates negative SHAP (protective), this confirms that working-poor context *suppresses* violations only in the subpopulation where it is below average—an unusual finding that likely reflects the specific service character of private utilities in mixed-income versus lower-income markets.

**PCD12** (Educational Residuals) Two-plateau step function with a high plateau (2.92) for  $PCD12 < -0.9$ , a steep decline between  $PCD12 \approx -0.9$  and  $-0.3$ , and a flat low plateau (2.33) for  $PCD12 > -0.2$ . The red dashed line at  $PCD12 = 0$  falls within the low plateau, confirming the transition completes in the negative

range. Range of effect: ~2.33 to ~2.92 (approximately –0.59 violation units—the largest marginal range of the two demographic PDPs shown). PCD12 (Educational Residuals) captures communities where educational attainment is anomalously high or low relative to what income and occupation levels would predict—a residual measure of educational structure. Low PCD12 (below-expected education given economic context) is associated with more violations by private utilities. The transition completing before zero means the zero-crossing has limited practical meaning for this variable: the decisive threshold is at  $PCD12 \approx -0.9$ . Communities with below-average educational residuals served by private utilities have approximately 0.6 more violations per decade than those with above-average educational residuals. This is consistent with the hypothesis that communities where educational attainment exceeds economic expectations have more civic engagement and technical literacy to monitor and challenge utility compliance behavior.

**PCF2 (Tax Impact)** Essentially a pure binary step function: flat very low plateau (2.13) for all  $PCF2 < 0$ , nearly vertical ascent at  $PCF2 \approx 0$ , flat very high plateau (4.47) for all  $PCF2 > 0.05$ . The red dashed line precisely marks the zero crossing, which is the decisive threshold. Range of effect: ~2.13 to ~4.47 (approximately +2.34 violation units—the **largest marginal range in this model**, exceeding even PCF1).

*Nonlinearity:* The most extreme step-function in this model and among the sharpest observed across all four models analyzed in this series. The entire transition occurs essentially at the  $PCF2 = 0$  boundary with essentially no gradient on either side. PCF2 (Tax Impact) is the **single most important predictor of private utility violations**, and its effect is enormous. Utilities where the tax component dominates the financial mix ( $PCF2 > 0$ ) average over 2.3 more violations per decade than utilities where operating expenses dominate ( $PCF2 < 0$ ). The zero boundary is the decisive threshold and the tick marks confirm real data density on both sides of it. This finding aligns with—but dramatically amplifies—the charges model's PCF2 effect: tax-dominant private utilities charge more *and* have far more violations. This suggests that PCF2 is not merely a financial strategy variable but a proxy for a specific type of private utility: those with high tax burdens relative to operating costs are likely large, complex, urban-adjacent investor-owned systems operating in high-tax jurisdictions (such as Pennsylvania), where both charge levels and regulatory complexity are elevated. The same utilities that can extract premium pricing appear to also generate compliance failures at a much higher rate—a compound disadvantage for their ratepayers.

**totexprevratio** (Total Expense-to-Revenue Ratio) Flat moderate plateau (2.67) for  $totexprevratio \approx 0.3$  to 0.75, then a step decline between approximately 0.75 and 1.1, reaching a flat low plateau (2.33) for all  $totexprevratio > 1.2$ . The red dashed line at zero falls to the left of all real data (tick marks begin near 0.3), so the zero crossing is an extrapolation marker rather than a transition point observed in the data. Range of effect: ~2.33 to ~2.67 (approximately –0.34 violation units—the smallest marginal range among the PDPs shown).

Private utilities with total expenses exceeding revenues (operating at an aggregate financial loss or deficit) have **fewer violations** than those running surpluses or balanced budgets. This is counterintuitive from a resource-capacity perspective but consistent with the charges model's PCF3 finding (negative margin → higher charges): cost-pressured utilities appear to be operating in different modes—either they are undergoing infrastructure rehabilitation funded by debt (high expenses, high compliance effort) or they are under heightened regulatory scrutiny that forces both cost expenditure and compliance. Utilities with low expense-to-revenue ratios (profitable, financially comfortable) may face less regulatory pressure and reduce compliance investment. The interaction dominance of *totexprevratio* (highest H-statistic) means this effect is strongly conditioned on financial context: the ratio's protective or harmful effect depends on what other financial characteristics the utility displays.

**PCD5 × taxpp**

**Key finding:** The joint PDP of PCD5 (Young Starter Families) and taxpp (taxes per customer) reveals a **three-zone structure** in which PCD5 creates distinct violation regimes that are essentially independent of taxpp, but taxpp provides a secondary, lower-amplitude modulation within each PCD5 regime.

- **Left zone ( $PCD5 < -1.2$ ):** Violations cluster at an intermediate level ( $\sim 2.59$ – $2.66$  across most of the zone), with a slight elevation at moderate taxpp levels. This corresponds to private utilities serving areas with very few young families—older, more established residential markets.
- **Central zone ( $-1.2 < PCD5 < +0.5$ ):** The lowest predicted violations ( $\sim 2.40$ – $2.53$ ) occur here, with the minimum concentrated in the middle PCD5 range at low taxpp. The narrow band at the very bottom of taxpp (actual data region) shows compressed contours between the 2.40 and 2.53 levels. This moderate-PCD5 zone represents the lowest-risk archetype for private utility violations.
- **Right zone ( $PCD5 > +1.0$ ):** Violations climb to the highest level in this interaction ( $\sim 2.72$ ), and the contours again become vertical—the elevated violation rate in high-PCD5 (young-family-dominated) areas is largely independent of taxpp.
- The taxpp axis (vertical) creates a minor gradient only at the very bottom of its range: low-taxpp utilities show slightly lower violations than high-taxpp utilities within each PCD5 regime, but the effect is small compared to the PCD5 boundary effects.

**Interpretation:** Young Starter Family concentration (PCD5) creates a U-shaped violation risk pattern: areas with the fewest and the most young families both face higher violations from private utilities than areas with moderate concentrations. The minimum-risk zone at moderate PCD5 coincides with neither extreme family-structure type—likely representing mixed suburban communities with stable demographics and adequate civic engagement. The right-zone elevation (high PCD5, more violations) is consistent with the idea that areas undergoing rapid demographic growth with new young families may have less-established infrastructure, newer service connections with teething problems, and weaker community advocacy capacity. The independence from taxpp at all but the lowest tax levels confirms the H-statistic finding that taxpp is a secondary interaction variable: PCD5 structure defines the violation regime, within which taxes per customer provide modest additional differentiation.

### E3.2.1 Synthesis: Conceptual Model of Private Utility Violation Drivers

#### *Tier 1: Financial Structure as the Primary Violation Signal*

**PCF2 (Tax Impact)** and **PCF1 (Financial Scale)** together constitute the dominant predictive signal for private utility violations, collectively accounting for the majority of model importance. Tax-dominant private utilities have dramatically more violations. Note that the relationship for PCF1 is compromised due to aggregation of PWSIDs that were acquired by larger private companies that had small, poor populations but very high violations. Their interaction with *totexprevratio* (the top H-statistic feature) means the financial structure of the utility is not merely additive: how tax burden and corporate scale interact with the expense-revenue balance shapes violation risk in a compounded way. This tier has no meaningful equivalent in the full-utility violations model, where water source type and state identity dominated. For private utilities, financial organization is the primary compliance risk factor.

#### *Tier 2: System Scale and Demographic Character*

**PCD6 (Large Stable Households)** is the leading demographic driver and operates in a positive direction—private utilities serving large stable household communities have more violations. **Population served** provides the only conventional scale-protection signal: smaller populations predict more violations, consistent with the full-utility finding. **PCD10 (The Working Poor)** and **PCD7 (Active Labor Force)** both

show protective (negative) effects, indicating that labor-market character of the service area modulates compliance risk in ways distinct from the affluence-driven patterns seen in the charges model.

### *Tier 3: Educational and Labor-Market Structure*

**PCD12 (Educational Residuals)** operates as the third demographic tier: areas with below-expected educational attainment relative to economic context have more violations. The interaction-ranking prominence of PCD12 (rank 4 in H-statistics despite rank 9 in importance) suggests it primarily moderates the financial-structure effects of Tier 1 rather than acting independently. **PCD5 (Young Starter Families)** creates a U-shaped violation landscape—moderate concentrations minimize risk, while extremes in either direction elevate it—with the 2D PDP confirming this operates largely independently of tax-per-customer levels.

### *Tier 4: Expense-Revenue Balance (Interaction Moderator)*

**totexprevratio** has a small marginal effect (a modest step near the break-even point) but the largest interaction H-statistic in the model. It functions primarily as a context-setter that conditions how financial structure and demographic features translate into violation risk. Utilities operating above their revenue (deficit spending) have modestly fewer violations, and this effect interacts with PCF1 and PCF2 to modulate the primary financial signal.

### *Tier 5: State Identity (Effectively Absent)*

**Pennsylvania** contributes a small positive SHAP effect—the only detectable state signal in this model. All other states are statistically indistinguishable from the national baseline after controlling for financial structure and demographics. This is a profound contrast with the full-utility violations model (Oklahoma-dominated) and indicates that the violation risk of private utilities is determined by what kind of financial and demographic entity they are, not where they are located.

## *E3.2.2 Conclusions*

**1. Tax-dominant private utilities are a high-violation, high-charge archetype requiring priority regulatory attention.** PCF2 is simultaneously the top predictor of charges and violations, with the same directional sign in both models. Ratepayers of tax-dominant investor-owned utilities face a double burden. Targeted compliance audits of all private utilities with positive PCF2 are warranted.

**2. PCD6 communities served by private utilities are underserved on compliance.** Large stable households—the archetypal suburban homeowner community—pay lower private utility charges but experience more violations. These communities may be insufficiently engaged in rate cases and compliance monitoring, creating an advocacy gap that private utilities are not filling. Community monitoring programs targeted to these service areas are recommended.

## Appendix F Dashboard for data and model exploration

A web based dashboard that allows the data and selected model results to be visualized is provided at <https://jo67893whu.github.io/MRCTI-dashboard/>.



The opening screen provides the choice of exploring a basic “**Utility Screen**”, “**Outliers and trends**”, “**Model Evidence**” and “**Data & Methods**”. Each is a clickable Menu that then allows further exploration. The Utility Screen allows the user to choose the data set and screen by a number of criteria.

The screenshot displays the 'Utility screening and geographic context' dashboard. At the top, it says 'NATIONAL COMMUNITY WATER SYSTEMS'. The main title is 'Utility screening and geographic context'. To the right of the title are 'Reset' and 'Export selection' buttons. Below the title is an 'Analytical scope' section with a text block explaining that EMMA is the national base layer and that 36 Private records without EPA identifiers use project-assigned RIK IDs. The main filtering area contains several dropdown menus: 'SEARCH' (PWSID, RIK ID, utility, company, or municipality), 'STATE' (All states), 'OWNERSHIP' (All ownership), 'SOURCE' (All sources), 'DATA LAYER' (All records), and 'SCREEN CONDITION' (No condition).

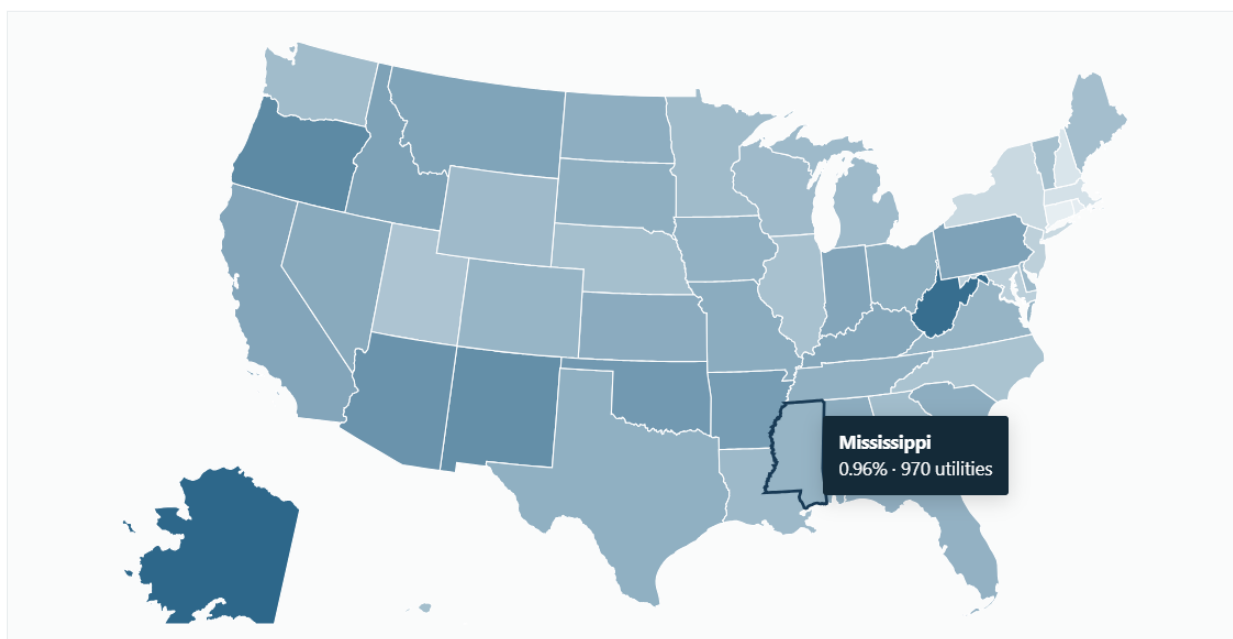
**Export Selection** (top right) allows the selected data to be exported as an Excel file.

**Search** allows a search by PWSID or private company name, or municipality, **State**, allows restriction to a specific state via a drop-down menu, or select all states (national) . **Ownership** allows filtering by an EPA Used ownership category via a drop-down menu. **Source** allows restriction of the water source via a drop-down menu. **Data Layer** allows the selection of the EMMA, Public or Private data sets via a drop-down menu. **Screen Condition** allows the selection of the variables of interest, as seen below.

The screenshot shows the 'SCREEN CONDITION' dropdown menu. The menu is open, showing the following options: 'No condition' (selected), 'Open health violation', 'Charge burden ≥ 2%', 'Any 10-year health violation', and 'Reported revenue/expense ratio < 1'.

## State-level view

Median charge burden



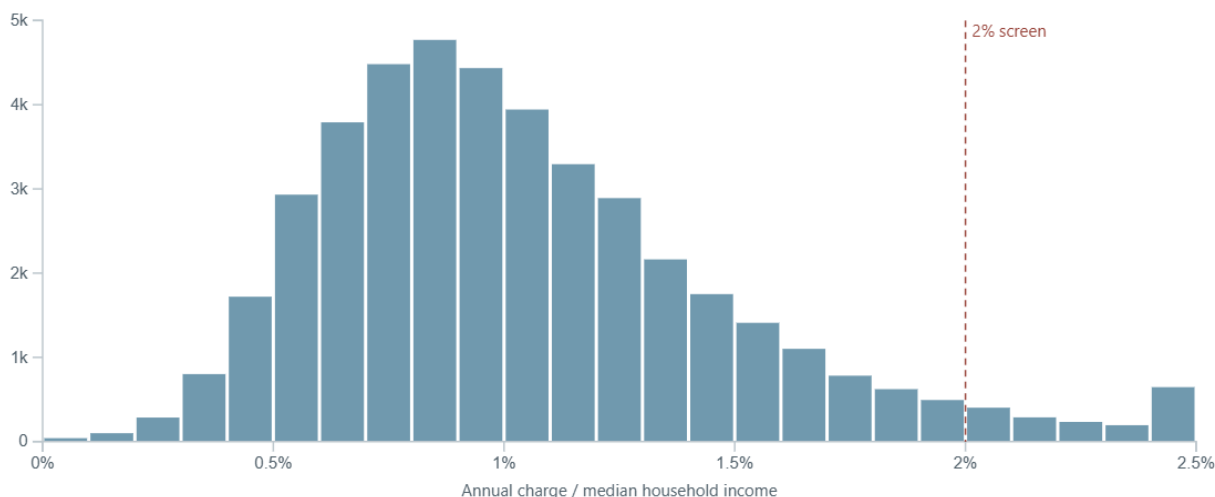
MAP MEASURE

- Median annual charge
- Utility count
- Median charge burden
- Median annual charge**
- 10-year health violations / 1,000
- Financial data coverage

In addition to the maps, a histogram of the selected variable is also provided, as well as a ranked screening table of the utilities for selected criteria as shown below. The entire data set can be downloaded for the selected data layer using Export Selection.

## DISTRIBUTION

### Annual charge burden



1,762 of 43,567 observed records meet or exceed the 2% screening threshold. Values above the 98th percentile are grouped in the final bin.

## SCREENING RECORDS

### Ranked screening table

43,567 records after filters

| RANK | UTILITY RECORD                                    | STATE | OWNERSHIP | POPULATION | RANK BY                               |  |
|------|---------------------------------------------------|-------|-----------|------------|---------------------------------------|--|
|      |                                                   |       |           |            | Charge burden                         |  |
| 1    | FOUNTAINVIEW SUBDIVISION<br>TX1010127             | TX    | Private   | 2,703      | Charge burden                         |  |
| 2    | PELICAN LAKE WATER PLANT<br>FL4505002             | FL    | Private   | 222        | Charge burden                         |  |
| 3    | CEDAR CREEK CABIN OWNERS ASSOCIATION<br>NM3558214 | NM    | Private   | 536        | 10-year health violations             |  |
| 4    | ALPINE VILLAGE W AND SD<br>NM3558014              | NM    | Private   | 294        | All-year total violations             |  |
| 5    | TALL PINES WATER ASSOCIATION<br>NM3580914         | NM    | Private   | 101        | Annual charge                         |  |
|      |                                                   |       |           |            | Population served                     |  |
|      |                                                   |       |           |            | Lowest reported revenue/expense ratio |  |

The second main menu item is **Outliers and Trends**. Here we can visualize the extreme cases for one or both response variables (Average charge and 10 year health violations) relative to what the model predicts. The idea is that the model uses all the information from the predictors for each pwsid and hence a charge that is much higher or much lower than is expected based on the model predictions that account for state, water source, demographics, utility scale etc is interesting to check at the utility level. In the illustration below we chose health violations that were in the top 5% of divergence from what is expected based on the model AND also charges that were in the top 5% relative to what was expected from the model. 39 Utilities are found that have super high charges and superhigh violations relative to expectation. Their position

relative to expectation is shown in the graph and clicking on a plotted point allows the specific utility information to be displayed. All utilities thus identified are also provided in rank order below.

## CASE-FINDING WORKSPACE

# Model outliers and violation trends

Separate model-relative anomalies from observed changes over time.

REPRODUCED

Report-era EMMA models

Health: N 43,531, CV R<sup>2</sup> 0.205 · Charge: N 43,531, CV R<sup>2</sup> 0.404 · five-fold out-of-fold predictions

### EMMA · MODEL-RELATIVE SCREEN

#### Out-of-fold residuals by PWSID

Residual = observed – cross-validated prediction. Large absolute residuals identify systems furthest from the fitted regression surface. Percentile flags are screening thresholds, not confidence intervals.

#### SEARCH

PWSID or available utility name

#### HEALTH RESIDUAL

Above model expect ▼

#### CHARGE RESIDUAL

Above model expect ▼

#### STATE

All states ▼

Matched utilities

39

Meeting both outcome conditions

Health extremes

39

Within the matched selection

Charge extremes

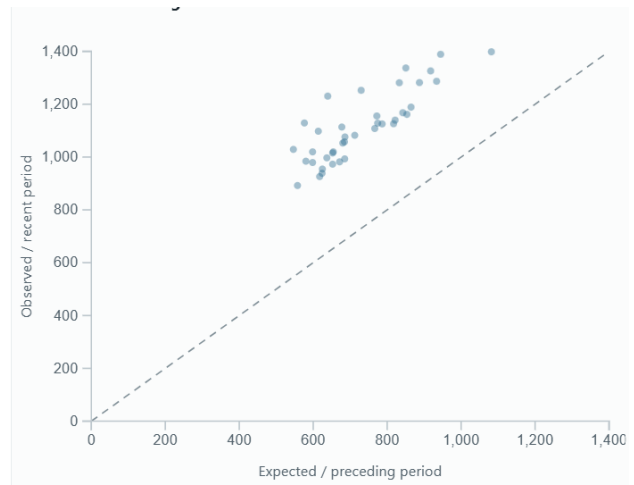
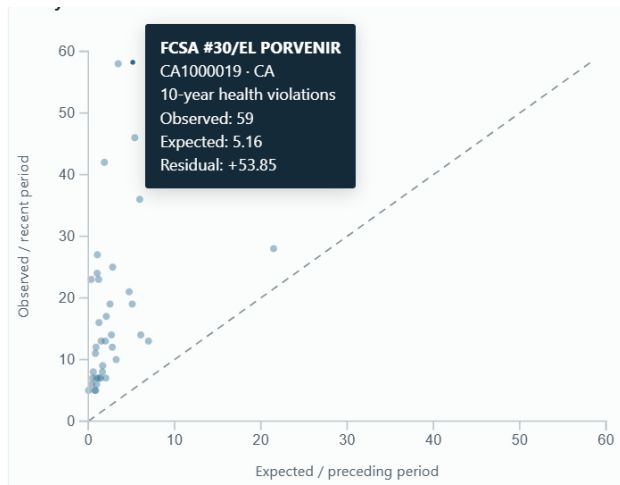
39

Within the matched selection

Dual extremes

39

Extreme for both outcomes

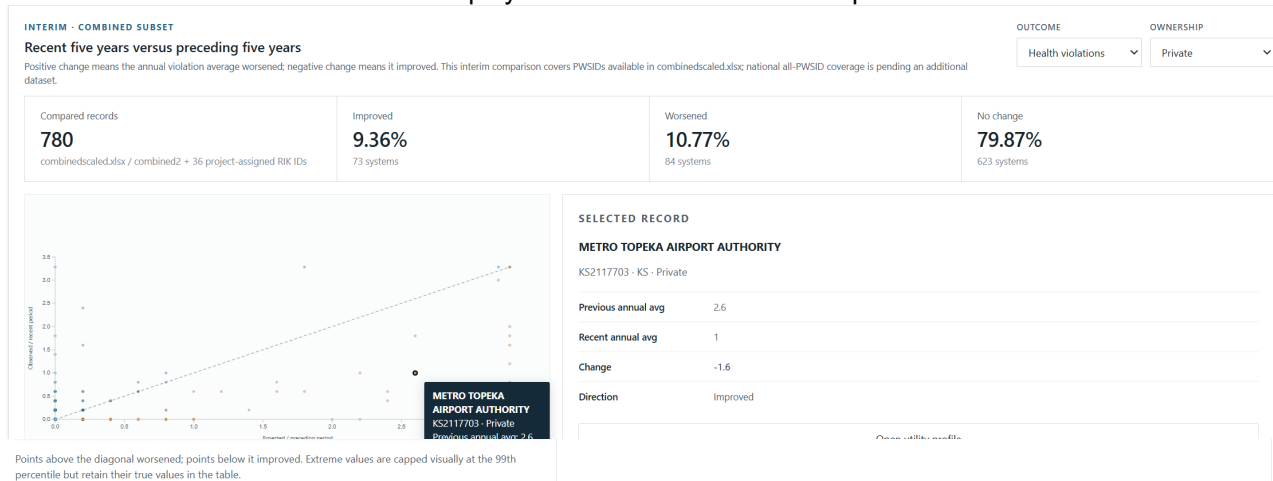


Both charts show the same jointly filtered utilities. The diagonal represents observed = expected; axes are capped at the 99th percentile for legibility.

| RANK | UTILITY RECORD                          | STATE | OWNERSHIP        | HEALTH RESIDUAL | HEALTH EXTREMENESS | CHARGE RESIDUAL | CHARGE EXTREMENESS |
|------|-----------------------------------------|-------|------------------|-----------------|--------------------|-----------------|--------------------|
| 1    | KOTZEBUE MUN. WATER SYSTEM<br>AK2340060 | AK    | Local government | +54.54          | 99.92%             | +\$353          | 97.12%             |
| 2    | FCSA #30/EL PORVENIR<br>CA1000019       | CA    | Local government | +53.85          | 99.91%             | +\$395          | 98.29%             |

The examples visible in the screenshot have substantially higher health violations and charges than expected, and their percentile is shown under extremeness. 99.92% means that only 0.08% were worse on health violations than this utility relative to what was expected, and 97.12% means that 2.88% of all utilities had higher charges relative expected.

The next comparison on the same page is of health violations or total violations, by ownership for the most recent 5 years vs the preceding 5 years. A screenshot is provided for private ownership, and it shows that most of the private utilities had no change in their violations. As in other plots, clicking on a plotted point allows that utilities information to be displayed. Ranked utilities are also provided.



**TREND OUTLIERS**

**Largest observed changes**

780 records after ownership filter

RANK BY: Largest absolute change

| RANK | UTILITY RECORD                      | STATE | OWNERSHIP | POPULATION  | 10-YEAR COUNT | PREVIOUS ANNUAL AVG | RECENT ANNUAL AVG | CHANGE | DIRECTION |
|------|-------------------------------------|-------|-----------|-------------|---------------|---------------------|-------------------|--------|-----------|
| 1    | EPCOR WATER ARIZONA INC.<br>RIK #16 | AZ    | Private   | 358,122,032 | 54            | 10.6                | 0.2               | -10.4  | Improved  |
| 2    | ATLANTA WATER SUPPLY<br>LA1127001   | LA    | Private   | 900         | 36            | 6                   | 1.2               | -4.8   | Improved  |

The next major menu item is **Model Evidence**. This provides visualizations of the Random Forest Models for the different data sets and response variables.

## REPORT MODEL EVIDENCE

# Model performance and comparability

A focused view of the report's two principal screening outcomes and the samples on which they were estimated.

### Comparability rule

Compare performance only within the same outcome. Sample size varies by analytical layer; encoded features count columns after categorical one-hot encoding, not original variables.

## CORE VALIDATION RESULTS

### Models supporting the first screening release

Average annual charge and 10-year health violations. Combined is excluded until its sample lineage is reconciled.

| OUTCOME                   | LAYER   | ANALYSIS N | ENCODED FEATURES | CV R <sup>2</sup> | CV RMSE | CV MAE  |
|---------------------------|---------|------------|------------------|-------------------|---------|---------|
| Average annual charge     | EMMA    | 43,531     | 82               | 0.404             | 152.073 | 116.501 |
| Average annual charge     | Public  | 8,007      | 88               | 0.317             | 153.067 | 120.469 |
| Average annual charge     | Private | 277        | 75               | 0.255             | 191.390 | 150.708 |
| 10-year health violations | EMMA    | 43,531     | 81               | 0.205             | 3.845   | 0.992   |
| 10-year health violations | Public  | 8,007      | 88               | 0.196             | 6.216   | 2.068   |
| 10-year health violations | Private | 277        | 75               | 0.044             | 8.656   | 3.610   |

## INTERPRETABILITY EVIDENCE

### What drives the fitted models

Final report-era random-forest outputs. Importance ranks predictive contribution; SHAP shows direction and dispersion of effects.

### DATASET

EMMA

Annual charge

10-year health violations

