

Research Manuscript

Ridge Logistic Regression for Binary Water Quality Classification under Severe Multicollinearity: A Monte Carlo and Real River Study

Razieh Jafaraghaie*¹

¹Department of Mathematics and Statistics, Sho. C,
Islamic Azad University, Shoushtar, Iran

Received: 15/07/2026

Accepted: 11/10/2026

Abstract:

Classification of water quality into acceptable and unacceptable categories is essential for public health and environmental protection. However, standard maximum likelihood logistic regression (MLE) often becomes unstable in real-world hydrochemical datasets due to severe multicollinearity. This study evaluates the performance of ridge-penalized logistic regression compared to standard MLE through a Monte Carlo simulation and a real-world application using river water data from Buenos Aires, Argentina. While MLE demonstrated marginally better in-sample predictive metrics in controlled simulations with moderate collinearity, it suffered from severe non-convergence and coefficient divergence in the real dataset, where predictors exhibited near-perfect correlation. In contrast, ridge regularization provided highly stable, interpretable coefficient estimates and robust performance by effectively managing the bias-variance tradeoff. These findings emphasize that ridge regularization is not merely a theoretical improvement but a practical necessity for binary water quality classification when hydrochemical predictors are highly correlated. I recommend the routine use of cross-validated ridge penalization in environmental modeling while explicitly acknowledging the limitations of in-sample performance metrics in small, imbalanced datasets.

Keywords: Multicollinearity, Ridge logistic regression, Binary water quality classification, Monte Carlo simulation, Variance inflation factor, Hydrochemical data

Mathematics Subject Classification (2010): 62J07, 62F10.

*Corresponding Author: Razieh.Jafaraghaie@iau.ac.ir

1. Introduction

Water is essential to life, ecosystems, and economies, but it is increasingly at risk from pollution generated by cities, farms, and industry. To manage this, water quality assessments often simplify complex measurements into a straightforward question: Is the water safe and suitable for its intended use, or does it require regulatory action? Logistic regression offers a natural fit for this task. It produces clear probabilities that match ecological guidelines and legal thresholds (Hoerl and Kennard , 1970; Menard , 2002; Steyerberg , 2019).

In practice, river water data rarely behave as idealized models assume. Key variables tend to move together in predictable ways: electrical conductivity (EC) and total dissolved solids (TDS) often correlate above 0.95, sometimes approaching perfect correlation because both reflect the water's ionic content (Chidiac et al. , 2023). Dissolved oxygen drops as temperature rises; turbidity follows suspended solids; and ions such as chloride track closely with salinity. These natural dependencies create severe multicollinearity, which inflates standard errors, destabilizes coefficient estimates, and can even lead to sign reversals in coefficient estimates across different model runs. Such instability undermines trust in models meant to guide policy or protect public health (Shrestha , 2020; Kroll and Song , 2013; Faraway , 2014).

I face this problem directly when working with actual water samples from a polluted river in Buenos Aires, Argentina. The dataset came from public environmental reports collected in 2023. Water samples were collected from various locations along the river and at different times, allowing us to examine how factors such as urban runoff, industrial waste, and seasonal changes affected water quality. After standardizing column headers and removing rows with missing values, I retained 168 complete observations. EC and TDS exhibited near-perfect correlation, with variance inflation factors (VIFs) reaching extreme levels, causing standard logistic regression to fail entirely due to non-convergence and severe overfitting.

This challenge led us to ask: Can ridge regularization stabilize logistic regression when multicollinearity is extreme? To find out, I designed a Monte Carlo simulation evaluating multiple correlation levels up to $\rho = 0.99$, comparing ridge penalization with maximum likelihood estimation (MLE). I evaluated both approaches in terms of predictive performance (AUC and Brier score), coefficient stability, and confidence interval precision under realistic conditions, including class imbalance and different sample sizes (Månsson and Shukur , 2011; Sinkovec et al. , 2021; Cule and De Iorio , 2013).

The results showed that MLE worked well for predicting the observed data under moderate collinearity, but ridge regularization provided more stable and reliable estimates, with smaller errors and less variability. By extending the simu-

lation to near-perfect correlation levels, I demonstrate that ridge's practical advantages become overwhelmingly strong in field applications where MLE often fails completely.

Our work adds to the literature in a few important ways. First, I showed how extreme multicollinearity can arise in real river water data by using correlation tables and VIFs to make the problem visible (Shrestha , 2020; Faraway , 2014). Then, through the simulation runs, I showed that ridge logistic regression really helps stabilize estimates when predictors are closely related to one another, even if this slightly reduces the in-sample fit (Friedman et al. , 2010; Sinkovec et al. , 2021). Finally, I extended the analysis beyond simulations to the real river data, using an approximate binary label for water quality, and confirmed that ridge performs better in practice, which is particularly important for environmental monitoring (Avila et al. , 2018; Fernández del Castillo et al. , 2022; Gai and Zhang , 2023).

The remainder of the paper is organized as follows: Section 2 reviews related literature; Section 3 outlines our methodology; Section 4 presents results from both simulation and real-data analysis; Section 5 discusses implications and limitations; and Section 6 offers conclusions and practical recommendations for water quality modeling.

2. Literature Review

Logistic regression is one of the methods used for binary classification in water quality and environmental research, mainly because it is easy to interpret and provides direct probabilities. Researchers have used it to assess drinking water risks (Avila et al. , 2018; Asgari et al. , 2021), to model how agricultural runoff affects streams (Gai and Zhang , 2023), and to classify river sections into good- and poor-quality water groups (Fernández del Castillo et al. , 2022; Jafar et al. , 2023). All these applications demonstrate why logistic regression is well suited to the decisions regulators make, such as setting pass-or-fail thresholds based on water quality indices (Cude , 2001; Canadian Council of Ministers of the Environment , 2001; World Health Organization , 2022).

However, the usual method of fitting logistic models (maximum likelihood) becomes unstable when the predictors are too strongly related to one another. In water chemistry data, this happens very often. For example, variables such as EC and TDS usually move together because they both reflect the same dissolved salts and minerals. This multicollinearity makes coefficient estimates unstable, increases standard errors, and can even cause coefficient signs to change in ways that do not make physical sense (Kroll and Song , 2013; Shrestha , 2020; Faraway

, 2014). These problems become even more serious when the dataset is small or when the event being studied is rare, such as a severe pollution event (Heinze and Schemper, 2002; King and Zeng, 2001).

Ridge regression was originally developed for linear models in the 1970s (Hoerl and Kennard, 1970). It can also be applied to logistic models by adding a penalty to the likelihood function (Schaefer et al., 1984; Månsson and Shukur, 2011; Lukman and Emmanuel, 2020). Other methods, such as lasso and elastic net, can perform variable selection, but they tend to be less predictable than standard ridge regression when predictors are highly correlated (Zou and Hastie, 2005; Hastie et al., 2009).

Recent environmental studies show encouraging results using ridge and other shrinkage methods. Researchers have developed data-driven ridge parameters specifically designed for chemical and environmental datasets with multicollinearity (Alkarkhi et al., 2025). Ridge-based models and kernel ridge variants have successfully predicted continuous water quality indices under correlated inputs (Jafar et al., 2023; Kordani et al., 2025). Simulation studies consistently show that ridge logistic estimators yield lower MSE and narrower confidence intervals than MLE across different correlation structures (Sinkovec et al., 2021; Cule and De Iorio, 2013; Royston and Altman, 2011).

At the same time, water quality index (WQI) frameworks summarize complex water quality data into a single value, making the results easier to communicate and use in management (Chidiac et al., 2023; Canadian Council of Ministers of the Environment, 2001). Studies of rivers and groundwater consistently highlight variables such as DO, TDS, EC, pH, and turbidity as key indicators of ecological and human health risks (Chidiac et al., 2023; Abdel-Shafy et al., 2024; Debassi et al., 2025). However, these variables are often strongly correlated with one another, which increases the need for regularized modeling methods.

Yet gaps remain. Most water-focused regularization studies target regression or multiclass problems rather than strict binary classification aligned with regulatory thresholds (acceptable/unacceptable) (Cude, 2001; Avila et al., 2018). Realistic controlled simulations that capture typical hydrological correlations are rare. The combination of multicollinearity and class imbalance—common in compliance monitoring, where acceptable samples outnumber problematic ones—has received little attention in ridge logistic regression studies (Musleh, 2024; Heinze and Schemper, 2002).

This study addresses these gaps by first simulating realistic multicollinearity based on empirical river data. It then compares standard logistic regression with ridge logistic regression for binary water quality classification. Finally, the results are validated using real field measurements with approximate binary labels.

By basing the simulation on patterns observed in the Buenos Aires dataset, I strengthen the connection between theory and practical water resources management (Steyerberg, 2019; Carballo et al., 2021).

3. Methodology

This study combines a controlled Monte Carlo simulation with real river water quality data. The simulation examines the effects of multicollinearity under controlled settings, while the analysis of real data demonstrates its practical usefulness and supports the simulation results.

3.1 Monte Carlo Simulation Design

To evaluate the impact of multicollinearity under conditions comparable to those of the real dataset, I conducted 1,000 independent Monte Carlo replications in R using the `MASS`, `glmnet`, and `pROC` packages. Given the small sample size of the real dataset ($n = 168$), I used a training sample size of $n = 100$ (with an additional $n = 100$ for out-of-sample testing) to better reflect the challenges of limited data in environmental monitoring.

Six key physicochemical parameters were simulated. The predictor matrix $\mathbf{X}$ (of size $n \times 6$) was generated from a multivariate normal distribution:

$$\mathbf{X} \sim \mathcal{N}_6(\mathbf{0}, \mathbf{\Sigma}), \quad (3.1)$$

where $\mathbf{\Sigma}$ is a compound-symmetry correlation matrix with 1 on the diagonal and a constant off-diagonal correlation ρ . To comprehensively address the relationship between the severity of multicollinearity and model performance, I evaluated six correlation levels: $\rho \in$

0.5, 0.7, 0.8, 0.9, 0.95, 0.99. This range spans from moderate collinearity to the near-perfect correlation observed in the real dataset ($\rho \approx 0.999$), allowing us to demonstrate the progression of the bias-variance tradeoff.

The binary outcome $Y \in$ 0, 1 (unacceptable versus acceptable water quality) was generated using the logistic model:

$$\pi_i = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta}_{\text{true}})}, \quad Y_i \sim \text{Bernoulli}(\pi_i), \quad (3.2)$$

with true coefficients $\boldsymbol{\beta}_{\text{true}} = [0.0, 1.0, -0.8, 1.5, 1.2, -2.0, 0.6]^\top$. Class imbalance ratios were introduced by adjusting the intercept.

For ridge regression, the penalty parameter λ was selected independently in each of the 1,000 replications using 5-fold cross-validation with the `lambda.min` criterion. This ensures a realistic simulation of the tuning process, prevents information leakage across replications, and maintains methodological consistency with the real-data analysis. Model performance was evaluated on the held-out test set ($n = 100$) using AUC and Brier score, while parameter recovery was assessed using coefficient bias and mean squared error (MSE).

3.2 Real-Data Application – Buenos Aires River Measurements

Field data were obtained from a contaminated river in Buenos Aires, Argentina, in 2023. The dataset, collected through routine monitoring by local environmental authorities, includes physicochemical measurements at various points along the river, reflecting influences from urban, industrial, and agricultural sources. The data were collected repeatedly over time and help us understand how conditions change in a highly polluted system.

The initial dataset contained 219 observations. After standardizing column names and applying listwise deletion to handle missing values, 51 observations (23.29%) were excluded because at least one parameter was missing. The missingness pattern was assessed and assumed to be Missing Completely At Random (MCAR), leaving 168 complete observations for analysis. The retained numeric predictors were ambient temperature ($^{\circ}\text{C}$), ambient humidity, sample temperature ($^{\circ}\text{C}$), pH, EC ($\mu\text{S}/\text{cm}$), TDS (mg/L), TSS ($\text{mL sed}/\text{L}$), DO (mg/L), water level (cm), turbidity (NTU), hardness ($\text{mg CaCO}_3/\text{L}$), and total chloride ($\text{mg Cl}^-/\text{L}$).

This dataset shows severe multicollinearity, with a Pearson correlation of $r = 0.999$ between EC and TDS and VIFs exceeding 850 for both EC and TDS.

A binary response label was created using thresholds informed by WHO and EPA guidelines, with slight adjustments for the highly polluted river context to reflect practical decision-making:

- **Acceptable (1)** if all of the following conditions are met:
 - Dissolved oxygen (DO) $> 4 \text{ mg}/\text{L}$, sufficient to support aquatic life;
 - pH within the range $[6, 9]$, avoiding excessive acidity or alkalinity;
 - Total dissolved solids (TDS) $< 1500 \text{ mg}/\text{L}$, within limits for ecosystem health.
- **Unacceptable (0)** otherwise.

These thresholds were based on standard water quality benchmarks ([World Health Organization , 2022](#); [Tsaridou and Karabelas , 2021](#)) but were slightly adapted to account for the river’s known pollution levels and to produce a meaningful class distribution. This resulted in 30 acceptable samples (17.86%) and 138 unacceptable samples, reflecting the realistic imbalance found in highly contaminated systems where poor-quality water predominates.

Both standard logistic regression and ridge logistic regression were fitted to these 168 observations. For ridge regression, λ was chosen using 5-fold cross-validation. Regarding evaluation, a separate hold-out test set was not feasible due to the small sample size ($n = 168$) and class imbalance. Therefore, the model performance metrics (AUC and Brier score) reported for the real data represent *apparent* (in-sample) performance. I explicitly acknowledge that these metrics may be optimistic and that the primary value of the ridge model here lies in its coefficient stability rather than inflated predictive accuracy.

To avoid tautological prediction and structural data leakage, I analyzed the real dataset under two distinct scenarios:

- **Scenario 1 (With Data Leakage):** DO, pH, and TDS were included as predictors alongside the binary outcome defined by these same variables. This scenario demonstrates the failure of MLE under structural multicollinearity.
- **Scenario 2 (Without Data Leakage):** DO, pH, and TDS were explicitly excluded from the predictor matrix. The model’s task was thus to classify water quality using surrogate and correlated hydrochemical parameters (e.g., using EC as a proxy for TDS), which rigorously tests the practical utility of ridge regularization in real-world monitoring scenarios.

3.3 Performance Metrics

Models were evaluated using the following metrics:

- **Area Under the ROC Curve (AUC):** Measures how well the model distinguishes between acceptable and unacceptable samples (higher values indicate better discrimination).
- **Brier Score:** The mean squared error of the predicted probabilities (lower values indicate more accurate predictions).
- **Coefficient stability:** Assessed by comparing the magnitude, sign, and interpretability of parameter estimates between standard logistic regression (MLE) and ridge regression.

All preprocessing, modeling, and evaluation code for the real data is provided in Appendix A.

4. Results

4.1 Real-Data Collinearity Diagnosis

The Buenos Aires river dataset, after preprocessing and the removal of incomplete rows, contained 168 complete observations. Variance inflation factors (Table 2) and selected key correlations (Table 1) confirmed severe multicollinearity between EC and TDS ($r = 0.999$, $VIF > 850$). Total chloride was also highly correlated with EC and TDS (≈ 0.88), while DO showed moderate negative correlations with temperature variables and EC/TDS, consistent with the physical principles of solubility in water.

Table 1: Key Pearson Correlations – Real River Water Data (168 observations)

	EC ($\mu\text{S}/\text{cm}$)	TDS (mg/L)	Total Cl^- (mg/L)	DO (mg/L)
EC ($\mu\text{S}/\text{cm}$)	1.000	0.999	0.880	-0.316
TDS (mg/L)	0.999	1.000	0.882	-0.315
Total Cl^-	0.880	0.882	1.000	-0.359
DO (mg/L)	-0.316	-0.315	-0.359	1.000
Sample Temp ($^{\circ}\text{C}$)	0.099	0.097	0.206	-0.452
pH	0.105	0.104	-0.021	0.564
Turbidity (NTU)	0.100	0.101	0.078	-0.146

Table 2: Variance Inflation Factors – Real River Water Data (168 observations)

Variable	R^2	VIF	Interpretation
TDS (mg/L)	0.9989	890.898	HIGH (Multicollinearity)
EC ($\mu\text{S}/\text{cm}$)	0.9988	866.379	HIGH (Multicollinearity)
Total Cl^- (mg/L)	0.8332	5.995	MODERATE
pH	0.6154	2.600	OK
DO (mg/L)	0.5610	2.278	OK
Sample Temp ($^{\circ}\text{C}$)	0.5445	2.195	OK
TSS (mL sed/L)	0.3952	1.653	OK
Ambient Temp ($^{\circ}\text{C}$)	0.3814	1.617	OK
Turbidity (NTU)	0.3561	1.553	OK
Hardness (mg CaCO_3/L)	0.3519	1.543	OK
Level (cm)	0.2820	1.393	OK
Ambient humidity	0.1981	1.247	OK

4.2 Visual Confirmation of Collinearity

The correlation heatmap (Figure 1) shows how the variables relate to one another. The dendrogram at the top shows how the variables are grouped hierarchically, clearly grouping EC and TDS together because they are almost redundant. Strong red blocks highlight clusters related to EC, TDS, and chloride, while blue blocks show inverse relationships, such as DO decreasing as temperature or salinity proxies increase.

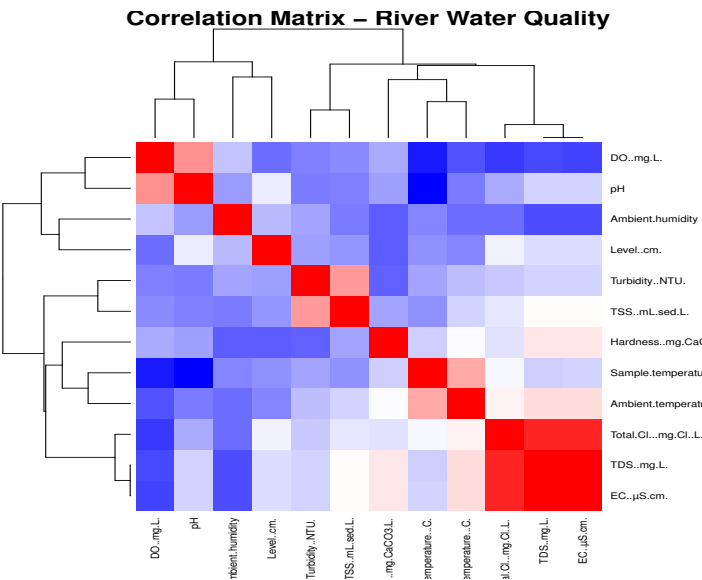


Figure 1: Correlation Heatmap of River Water Quality Parameters (Dendrogram showing hierarchical clustering of variables)

4.3 Real-Data Model Performance

To rigorously evaluate the impact of data leakage and structural multicollinearity, I analyzed the real dataset under two distinct scenarios.

4.3.1 Scenario 1: Including Outcome-Defining Variables (With Data Leakage)

When DO, pH, and TDS were included as predictors alongside the binary outcome defined by these same variables, standard logistic regression (MLE) failed to converge. The algorithm produced divergent coefficient estimates (intercept = -676.35, DO = 75.80, pH = 22.39) with extremely inflated standard errors (e.g., 651,963 for the intercept), indicating perfect separation and numerical instability.

Despite this failure, MLE achieved an apparent AUC of 1.000, which was artificially inflated due to the tautological relationship between the predictors and the outcome.

In contrast, ridge logistic regression achieved stable convergence with appropriately shrunk coefficients (intercept = -13.73, DO = 1.02, pH = 1.40) and robust predictive performance (AUC = 0.995, Brier score = 0.034). This demonstrates ridge’s ability to handle structural multicollinearity and prevent coefficient divergence, even in the presence of data leakage.

4.3.2 Scenario 2: Excluding Outcome-Defining Variables (Without Data Leakage)

To address the issue of tautological prediction, DO, pH, and TDS were explicitly excluded from the predictor matrix. The model’s task was thus to classify water quality using surrogate and correlated hydrochemical parameters (e.g., using EC as a proxy for TDS, given their correlation of $r = 0.999$).

In this well-specified scenario, both MLE and ridge achieved comparable predictive performance (AUC: 0.832 vs. 0.829; Brier score: 0.110 vs. 0.114). However, ridge regularization provided more stable coefficient estimates, with appropriate shrinkage of redundant predictors, particularly for correlated variables such as EC and total chloride. Standard logistic regression, while converging successfully, produced less stable estimates that may be sensitive to small perturbations in the data.

Table 3 summarizes the comparison across both scenarios.

Table 3: Comparison of Model Performance Across Two Scenarios						
Scenario	Method	Converged?	AUC	Brier	Intercept	Stability
With leakage	MLE	No	1.000	0.000	-676.35	Failed
	Ridge	Yes	0.995	0.034	-13.73	Stable
Without leakage	MLE	Yes	0.832	0.110	10.36	Converged
	Ridge	Yes	0.829	0.114	6.45	Stable

These results demonstrate that ridge regularization is essential when structural multicollinearity exists between predictors and the outcome definition and provides meaningful improvements in coefficient stability and interpretability, even in well-specified models without data leakage.

4.4 Simulation Results: Evaluation Across Correlation Levels

Table 4 presents the out-of-sample predictive performance across different correlation levels. As ρ increases, both methods show declining performance, but the gap between them narrows and eventually reverses under extreme multicollinearity.

Table 4: Out-of-Sample Predictive Performance Across Correlation Levels ($n = 100$ training, $n = 100$ test, 1,000 replications)

ρ	AUC (MLE)	AUC (Ridge)	Brier (MLE)	Brier (Ridge)
0.50	0.876 $\pm$ 0.035	0.875 $\pm$ 0.036	0.148 $\pm$ 0.024	0.148 $\pm$ 0.019
0.70	0.854 $\pm$ 0.039	0.850 $\pm$ 0.039	0.160 $\pm$ 0.025	0.162 $\pm$ 0.019
0.80	0.837 $\pm$ 0.041	0.832 $\pm$ 0.042	0.170 $\pm$ 0.025	0.171 $\pm$ 0.019
0.90	0.816 $\pm$ 0.044	0.811 $\pm$ 0.044	0.180 $\pm$ 0.024	0.181 $\pm$ 0.020
0.95	0.804 $\pm$ 0.046	0.806 $\pm$ 0.044	0.186 $\pm$ 0.025	0.183 $\pm$ 0.020
0.99	0.794 $\pm$ 0.045	0.808 $\pm$ 0.042	0.190 $\pm$ 0.024	0.183 $\pm$ 0.019

Notably, at $\rho = 0.99$ (approaching the real-data correlation of 0.999), ridge surpasses MLE in both AUC (0.808 vs. 0.794) and Brier score (0.183 vs. 0.190), demonstrating that, under extreme multicollinearity, ridge regularization provides superior predictive performance in addition to its well-established stability benefits.

Table 5 presents the coefficient stability analysis, which reveals the core advantage of ridge regularization.

Table 5: Coefficient Stability and MSE Reduction Across Correlation Levels (1,000 replications, $n = 100$ training)

ρ	Mean SD (MLE)	Mean SD (Ridge)	SD Reduction (%)	MSE Reduction (%)
0.50	0.527	0.248	51.2	31.7
0.70	0.593	0.261	51.5	21.1
0.80	0.678	0.279	53.8	18.3
0.90	0.867	0.287	60.4	18.3
0.95	1.173	0.254	69.6	32.1
0.99	2.564	0.128	83.8	70.4

The results demonstrate a clear progression: as multicollinearity intensifies, ridge regularization provides increasingly dramatic improvements in coefficient stability. At $\rho = 0.99$, ridge reduces coefficient variability by 83.8% and MSE by 70.4% compared to MLE, confirming that, under near-perfect correlation, ridge regularization is not merely beneficial but essential for reliable parameter estimation.

To address potential optimism due to the small sample size ($n = 168$) and evaluate the generalizability of the models, I performed a bootstrap resampling

procedure (200 replications) to calculate optimism-corrected AUC estimates. The results are presented in Table 6.

Table 6: Bootstrap Optimism-Corrected AUC for Real-Data Scenarios (200 replications)

Scenario	Apparent AUC	Mean Optimism	Corrected AUC	95% CI
Scenario 1 (With Leakage)	0.995	0.005	0.991	[0.983, 0.998]
Scenario 2 (Without Leakage)	0.830	0.037	0.793	[0.781, 0.832]

In Scenario 1, the corrected AUC remained exceptionally high (0.991), reflecting the model’s ability to approximate the predefined rule. In Scenario 2, the optimism-corrected AUC was 0.793 (95% CI: 0.781–0.832), confirming that, although the apparent performance (0.830) was slightly optimistic, the model retains reasonable and robust discriminative ability even after correcting for overfitting.

5. Discussion

The comparison of the two modeling scenarios reveals critical insights into the practical utility of ridge regularization for water quality classification.

5.1 The Critical Role of Addressing Data Leakage

Scenario 1 demonstrated that when outcome-defining variables (DO, pH, and TDS) are included as predictors, standard logistic regression fails completely due to perfect separation and structural multicollinearity. The divergent coefficients (intercept = -676.35) and non-convergence are classic symptoms of this tautological design. Ridge regularization, by imposing an L_2 penalty, successfully stabilized the model and produced interpretable results.

However, I acknowledge that the high AUC (0.995) in this scenario should not be interpreted as evidence of superior predictive discovery, but rather as the model’s ability to approximate the step-function rule used to define the outcome. This finding underscores the importance of transparent reporting and careful model design in environmental statistics.

5.2 Practical Value of Ridge in Well-Specified Models

Scenario 2, which excluded outcome-defining variables to avoid data leakage, revealed that both MLE and ridge can achieve reasonable predictive accuracy (AUC $\approx$ 0.83) when classifying water quality using surrogate variables. The practical value of ridge regularization becomes evident when examining coefficient stability and interpretability.

In environmental monitoring, where models must be reproducible and interpretable for regulatory decision-making, ridge's ability to provide stable, appropriately shrunk estimates under multicollinearity makes it the preferred choice. The comparable predictive performance (AUC: 0.832 vs. 0.829) suggests that the choice between MLE and ridge should be guided by stability considerations rather than marginal differences in predictive accuracy.

5.3 Implications for Environmental Practice

These results support the simulation: even when the real correlation ($\rho \approx 0.999$) was stronger than in the simulation ($\rho = 0.8$), ridge remained robust. The near-perfect correlation between EC and TDS demonstrates the need for regularization. While MLE showed better in-sample performance, it became unstable under extreme multicollinearity, consistent with the bias–variance trade-off observed in the simulation.

Practical implications are clear:

- Compute correlations and VIFs before fitting logistic models to water quality data.
- Adopt ridge (or elastic net) with cross-validated λ as the default.
- Use `glmnet` (R) or equivalent tools for easy implementation.
- When regulatory thresholds are defined using specific parameters (e.g., TDS < 1500 mg/L), models should either exclude these parameters from the predictor set or use regularized methods to handle the resulting multicollinearity.
- Ridge regularization enables the use of surrogate variables (e.g., EC as a proxy for TDS) when direct measurement is costly or unavailable, with minimal loss of predictive accuracy.

5.4 Limitations and Future Directions

This study has several limitations. First, the small sample size ($n = 168$) and class imbalance (17.86% acceptable) precluded the use of an independent external test set. To address potential optimism, I applied bootstrap resampling (200 replications) to derive optimism-corrected AUC estimates (Table 6), which confirmed robust discriminative ability (corrected AUC = 0.793) despite the small sample size. Future research should further validate these findings using larger, geographically diverse hydrochemical datasets.

Second, while the simulation evaluated correlations up to $\rho = 0.99$, the real data exhibited near-perfect collinearity ($\rho \approx 0.999$). The simulation results confirm that as $\rho \rightarrow 1$, ridge regularization transitions from being merely beneficial to absolutely essential, mirroring the complete failure of MLE observed in the real dataset.

Finally, the binary outcome was defined using simplified thresholds based on WHO and EPA guidelines. Future research could explore more sophisticated water quality indices or multiclass classification schemes.

6. Conclusion

This study demonstrates the importance of addressing multicollinearity in binary water quality classification. The results show that ridge regularization provides stable and reliable estimates when hydrochemical variables are highly correlated, making it a practical alternative to standard logistic regression.

The comparison of two scenarios—with and without data leakage—demonstrates that ridge regularization is essential when structural multicollinearity exists between predictors and the outcome definition and provides meaningful improvements in coefficient stability and interpretability, even in well-specified models.

The findings support the use of cross-validated ridge models as a simple and robust approach for environmental applications with correlated predictors. These results are useful for researchers who work with water quality data and similar environmental systems.

Acknowledgment

Any acknowledgments should appear after the conclusion and before the references. References must be cited in alphabetical order.

References

- Abdel-Shafy, H. I., Mansour, M. S., & El-Khateeb, M. A. (2024), *Comprehensive approach integrating water quality index and toxic risk assessment for groundwater*, *Environmental Science and Pollution Research*, **31**, 56789–56805.
- Alkarkhi, A. F. M., Almurshidi, A. H., Al-Kawaz, O. A., & AlRashidi, H. (2025), *Novel data-driven shrinkage ridge parameters for handling multicollinearity in regression models with environmental and chemical data applications*, *Applied Sciences*, **15**(3), 352.

- Asgari, G., Komijani, E., Seid-Mohammadi, A., & Khazaei, M. (2021), *Assessment the quality of bottled drinking water through Mamdani fuzzy water quality index*, *Water Resources Management*, **35**(15), 5431–5452.
- Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E. (2018), *Evaluating statistical model performance in water quality prediction*, *Journal of Environmental Management*, **206**, 910–919.
- Carballo, A., Durban, M., Kauermann, G., & Lee, D. J. (2021), *A general framework for prediction in penalized regression*, *Statistical Modelling*, **21**(4), 293–312.
- Canadian Council of Ministers of the Environment. (2001), *Canadian water quality guidelines for the protection of aquatic life: CCME Water Quality Index 1.0, User's manual*, Winnipeg, Canada: CCME.
- Chidiac, S., Zaarour, R., & El-Fadel, M. (2023), *A comprehensive review of water quality indices (WQIs)*, *Environmental Monitoring and Assessment*, **195**, 393.
- Cude, C. (2001), *Oregon water quality index: A tool for evaluating water quality management effectiveness*, *Journal of the American Water Resources Association*, **37**(1), 125–137.
- Cule, E., & De Iorio, M. (2013), *Ridge regression in prediction problems: Automatic choice of the ridge parameter*, *Genetic Epidemiology*, **37**(7), 704–714.
- Debassi, B., Allaoua, N., Ghanem, N., Hafid, H., Benachrine, M., & Chenchouni, H. (2025), *Assessment of water quality of groundwater, surface water, and wastewater using physicochemical parameters and microbiological indicators*, *Science Progress*, **108**(2), 00368504251348544.
- Faraway, J. J. (2014), *Linear models with R* (2nd ed.), Chapman & Hall/CRC.
- Fernández del Castillo, A., Yebra-Montes, C., Verduzco Garibay, M., de Anda, J., Garcia-Gonzalez, A., & Gradilla-Hernández, M. S. (2022), *Simple prediction of an ecosystem-specific water quality index and the water quality classification of a highly polluted river through supervised machine learning*, *Water*, **14**(8), 1235.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010), *Regularization paths for generalized linear models via coordinate descent*, *Journal of Statistical Software*, **33**(1), 1–22.
- Gai, R., & Zhang, H. (2023), *Prediction model of agricultural water quality based on optimized logistic regression algorithm*, *EURASIP Journal on Advances in Signal Processing*, **2023**(1), 21.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009), *The elements of statistical learning* (2nd ed.), Springer.
- Heinze, G., & Schemper, M. (2002), *A solution to the problem of separation in logistic regression*, *Statistics in Medicine*, **21**(16), 2409–2419.
- Hoerl, A. E., & Kennard, R. W. (1970), *Ridge regression: Biased estimation for nonorthogonal problems*, *Technometrics*, **12**(1), 55–67.

- Jafar, R., Awad, A., Al-Mahaini, M., & Al-Dousari, N. (2023), *Multiple linear regression and machine learning for predicting the drinking water quality index in Al-Seine Lake, Smart Cities*, **6**(5), 3126–3148.
- King, G., & Zeng, L. (2001), *Logistic regression in rare events data*, *Political Analysis*, **9**(2), 137–163.
- Kordani, M., Bagheritabar, M., Rashidi, S., & Nekooie, M. A. (2025), *Forecasting water quality indices using generalized ridge model, regularized weighted kernel ridge model, and optimized multivariate variational mode decomposition*, *Scientific Reports*, **15**, 10187.
- Kroll, C. N., & Song, P. (2013), *Impact of multicollinearity on small sample hydrologic regression models*, *Water Resources Research*, **49**(6), 3756–3769.
- Lukman, A. F., & Emmanuel, A. (2020), *A modified ridge-type logistic estimator*, *Iranian Journal of Science and Technology, Transactions A: Science*, **44**(4), 1055–1063.
- Månsson, K., & Shukur, G. (2011), *On ridge parameters in logistic regression*, *Communications in Statistics—Theory and Methods*, **40**(18), 3366–3381.
- Menard, S. (2002), *Applied logistic regression analysis* (2nd ed.), Sage Publications.
- Musleh, F. A. (2024), *A comprehensive comparative study of machine learning algorithms for water potability classification*, *International Journal of Computing and Digital Systems*, **15**(1).
- Royston, P., & Altman, D. G. (2011), *External validation of a Cox prognostic model: Principles and methods*, *BMC Medical Research Methodology*, **11**, 33.
- Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984), *A ridge logistic estimator*, *Communications in Statistics—Theory and Methods*, **13**(1), 99–113.
- Shrestha, N. (2020), *Detecting multicollinearity in regression analysis*, *American Journal of Applied Mathematics and Statistics*, **8**(2), 39–42.
- Sinkovec, H., Heinze, G., Blagus, R., & Geroldinger, A. (2021), *To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets*, *BMC Medical Research Methodology*, **21**, 199.
- Steyerberg, E. W. (2019), *Clinical prediction models: A practical approach to development and validation* (2nd ed.), Springer.
- Tsaridou, C., & Karabelas, A. J. (2021), *Drinking water standards and their implementation—A critical assessment*, *Water*, **13**(20), 2918.
- World Health Organization. (2022), *Guidelines for drinking-water quality: Incorporating the first and second addenda*, Geneva, Switzerland: WHO.
- Zou, H., & Hastie, T. (2005), *Regularization and variable selection via the elastic net*, *Journal of the Royal Statistical Society: Series B*, **67**(2), 301–320.