Predicting the *E. coli* Concentration in Water Bodies

Bryce Clement and **Ryan Wu**
{brclement,rywu}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

Monitoring *Escherichia coli* concentrations in recreational water is crucial for public health and safety. In this project, we applied regression models to U.S. Geological Survey (USGS) datasets from Beach 6 and Huntington Beach, two Great Lakes sites, to predict *E. coli* concentrations from environmental factors. For Beach 6, these factors included turbidity, relative humidity, water temperature, number of birds, lake level change, wind speed, and rainfall. For Huntington Beach, these factors included lake temperature, lake turbidity, wave height, lake level change, and rainfall. Our best models outperformed or matched the published USGS baselines. Although our models achieved statistical success, we believe that the real-life application of the models should be taken with caution.

1 Introduction

Escherichia coli is a common bacterium. While commensal strains are generally harmless, elevated *E. coli* in natural waters is a public health concern and often indicates other microbial hazards. Therefore, it is important to understand and monitor *E. coli* concentrations to assess the water quality better.

Developing a fast and efficient quantification of *E. coli* concentration has been a crucial goal when monitoring water quality for water bodies. Previous studies have developed successful machine learning models using RGB imagery to predict *E. coli* concentration (Seok Min Hong 2024). It gave us the insight that using certain parameters as simple as color components could predict the *E. coli* concentration.

In this study, we used the Sciencebase dataset that focuses on water quality data at recreational sites in Ohio and Pennsylvania as part of the Great Lakes. The dataset contains categories such as turbidity, humidity, temperature, etc. to correlate corresponding *E. coli* concentrations in two water bodies: Beach 6 and Huntington.

We primarily evaluate regularized linear models, adding a degree-2 polynomial variant to capture interactions where warranted. The models are to predict the *E. coli* concentration using given parameters related to water quality and environmental status.

In the remainder of the paper, we introduce the background information of the data contained in our datasets, our experimental design using scikit-learn packages and regression models, and our results regarding the accuracy of models. Eventually, we will also discuss the broader impact of this project and our reflections.

2 Background

The original data from Beach 6 and Huntington contain different parameters with different pre-processing requirements as described on the website. To make sure our model aligns with the existing model for a more direct comparison, we decided to operate pre-processing according to the documentation. Below are the attributes of both datasets; pre-processing will be mentioned if it is required and performed by us from the original data.

Beach 6

1. ECOLI_LOG10
   - Description: concentration of *E. coli* recorded
   - Already $\log_{10}$ transformed
2. TURB_NTRU
   - Description: turbidity value
   - Pre-processing: $\log_{10}$ transform
3. RHUM_PCT
   - Description: relative humidity
4. WTEMP_CEL
   - Description: water temperature in degree Celsius
5. BIRDS_NO
   - Description: number of birds spotted around and in the bathing area
6. CHANGELL_FT
   - Description: lake level change for the past 24 hours
7. AirportWindSpInst_mph
   - Description: wind speed at 8am
8. AirportRain48W_in
   - Description: cumulative rainfall for the past 48 hours
   - Pre-processing: square-root transform

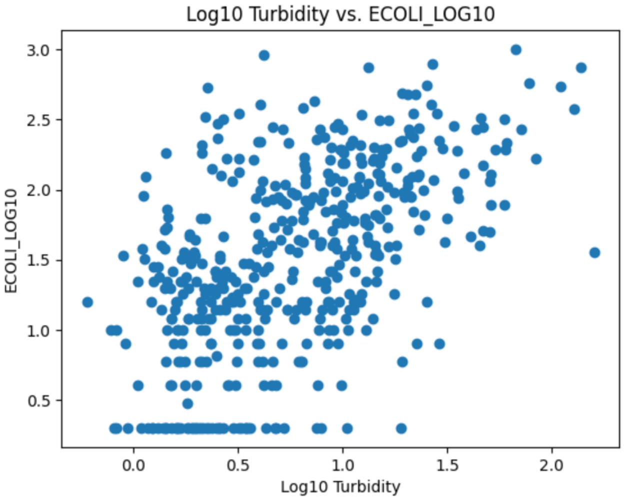

Figure 1: Correlation between $\log_{10}$ turbidity and $\log_{10}$ *E.coli* concentration.

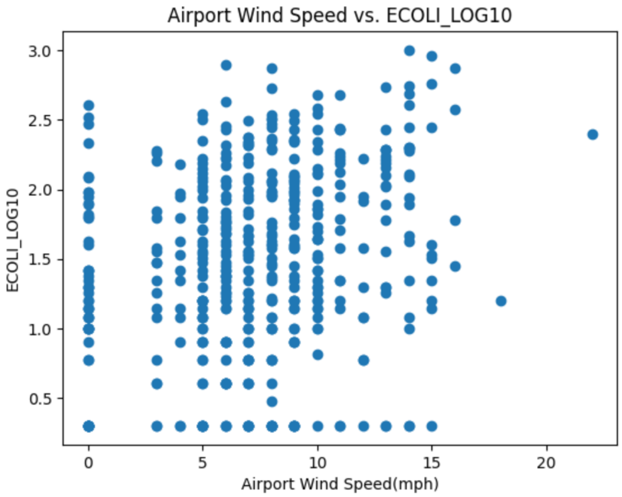

Figure 2: Correlation between airport wind speed and $\log_{10}$ *E. coli* concentration.

Huntington

1. EcoliAve_CFU
   - Description: *E. coli* concentration
   - Pre-processing: $\log_{10}$ transform
2. Lake_Temp_C
   - Description: lake temperature in degrees Celsius
3. Lake_Turb_NTRU
   - Description: turbidity of the water body
   - Pre-processing: $\log_{10}$ transform
4. WaveHt_Ft
   - Description: wave height at the time the E.coli sample was collected
   - Pre-processing: square-root transform
5. LL_PreDay
   - Description: lake level change for the past 24 hours
6. AirportRain48W_in
   - Description: cumulative rainfall for the past 24 hours
   - Pre-processing: square-root transform

Data Exploration

Before diving into building our model, we explored the dataset and plotted correlation graphs to understand the relationships between $\log_{10}$ *E. coli* concentration and key features. For Beach6, we found that turbidity shows a strong linear association (Figure 1). We also noted that discrete data points, such as airport instantaneous wind speed (AirportWindSpInst_mph), can introduce step-like patterns that complicate linear fits and their visual interpretation (Figure 2).

3 Experiments

Training Strategy & Technique

Both datasets were split using the 80-20 train/test split. Then we ran a 5-fold cross-validation on the training set only. We tuned the hyperparameters of the model with GridSearchCV. The hyperparameters we tuned were alpha for Ridge and Lasso models, the penalty, alpha, and l1-ratio for SGD models, and the degree and alpha for PolyRidge models. The hyperparameters that maximized CV R-squared were selected. This approach provides a low leakage estimate of generalization and guards against arbitrary hyperparameter selection. After selecting the best model using cross-validation, we refit it on the entire training set and evaluated it on the untouched test set, reporting R-squared and RMSE for an unbiased comparison with USGS baselines.

Model 1: SGDRegressor

Initially, we picked SGDRegressor as our first linear model to test. It uses the strategy of stochastic gradient descent to take each sample one at a time when measuring the gradient of loss. We set up using the L2 penalty by default to start with a maximum iteration of 1000.

Model 2: Ridge

We then switched from stochastic to batch-based models. This is because we wanted to see if the use of a batch-based model could produce a smoother convergence with a more accurate gradient. We first used ridge regression to serve as a direct comparison with stochastic gradient descent.

Model 3: LASSO

The purpose of trying LASSO was to test the change when switching from the L2 to the L1 penalty. We expect the model to generate coefficients that reflect the relevance of different parameters.

Model 4: Polynomial Ridge

We decided to move on from a linear regression model to a polynomial model because we wanted to explore what the performance would be if trained with higher-degree features. We used `PolynomialFeatures` from the `sklearn.preprocessing` to achieve it.

Statistical Analysis

We used the Python package `statsmodels` to help us assess the quality of our model. We did R-squared and p-value analysis to measure the fitness and accuracy of each coefficient of hyperparameters across different models.

Differential Analysis

We performed a differential analysis to sort out the top ten outliers with the greatest divergence between the predicted and actual values in our final coding files. We achieved this by pooling the predicted and actual values, calculating their differences, and sorting the absolute values in descending order. Eventually, we were able to generate a table of the top 10 outliers with their hyperparameter values.

4 Results

Our final results and code are in `b6_model_final.ipynb` and `huntington_model_final.ipynb`. We evaluated the quality of fitting across all our models using the R-squared value and root mean square error (RMSE) analyses.

Beach 6

For the Beach 6 dataset, we found that all four models we generated have higher R-squared values while maintaining lower RMSE values than the published model (Figure 3). This result surprised us, although we realized that our fine-tuning of coefficients using cross-validation and grid search could contribute to better performance. Having lower scores on training data, yet higher scores on testing data, indicates the absence of overfitting in our models.

Looking into each model for the Beach 6 dataset, we found that the performance of Ridge and PolyRidge was the same. We then plotted the graphs of actual values versus predicted values for both models (Figure 5, 6). We observed a consistent pattern of prediction versus actuality. This was due to the preference of hyper-parameter dimension by the grid search. We found that in our code, grid search selected degree = 1, which reduces to a linear model, leading to the identical performance by Ridge and PolyRidge.

The result of our differential analysis suggested that the prediction could be strongly affected by the hyperparameter `BIRDS_NUM`. The huge swing between discrete numbers could potentially contribute huge variance during prediction. The p-value of the `BIRDS_NUM` coefficient also suggested that, compared to other coefficients, it is more prone to generate variant results.

Huntington

For the Huntington dataset, our models generally performed as well as the published one (Figure 4). The PolyRidge model wins with an exceptional 0.56590 R-squared value and 0.4336 RMSE. It indicates that the connections between hyperparameters are of significance to the prediction of *E. coli* concentration in Huntington.

Model	Train R^2 Mean	Train RMSE Mean	Test R^2	Test RMSE
Published	N/A	N/A	0.47697	0.4841
Ridge	0.42921	0.5032	0.48872	0.4431
PolyRidge	0.42921	0.5032	0.48872	0.4431
SGD	0.42833	0.5029	0.48851	0.4432
LASSO	0.42817	0.5033	0.48912	0.4429

Figure 3: Model performance on the Beach 6 dataset

Model	Train R^2 Mean	Train RMSE Mean	Test R^2	Test RMSE
Published	N/A	N/A	0.54985	0.4432
PolyRidge	0.54241	0.4423	0.56590	0.4336
Ridge	0.53913	0.4438	0.54807	0.4424
LASSO	0.53904	0.4438	0.54774	0.4426
SGD	0.53856	0.4440	0.54826	0.4423

Figure 4: Model performance on the Huntington dataset

We also plotted the predicted versus actual values for the PolyRidge and Ridge models to see the differences. We observed that although the patterns are consistent when on the lower end ($\log_{10}$ *E. coli* concentration ≤ 2.0), the PolyRidge model produces more converged data on the higher end ($\log_{10}$ *E. coli* concentration > 2.0). This suggested that in order to predict higher values better, a higher degree of hyperparameters is necessary. The differential analysis unravels the effect of the hyperparameter `WaveHt_Ft` on the prediction. We found multiple zero values among the top ten outliers. Since the model is using polynomial regression, any combination of hyperparameters associated with `WaveHt_Ft` would be zero. This could cause a major fluctuation between the predicted and actual values.

5 Broader Impacts

Our models predict the *E. coli* concentration in water bodies from the Great Lakes region in Ohio and Pennsylvania. Based on data points related to such as water and weather conditions, we produced models that have similar or even better performances than the published model, indicating the possibility of generating better prediction using ours. However, a few concerns still remain in our mind at the end of this project.

The first concern is the generality of our model. Since the data was collected near two beaches that are considered as on-shore or near-shore areas, the models are limited by the data because it could only be applicable to similar environments. Furthermore, the prediction should not be reliable for predicting water quality in regions that are not similar as the Great Lakes region, which has a unique and complex system of environmental components. We believe that although the results of our models look good on the paper, the practicality of actual real-life use should be concerned before application.

The second is about the general data-driven prediction of water quality. We believe that it should be cautious to take

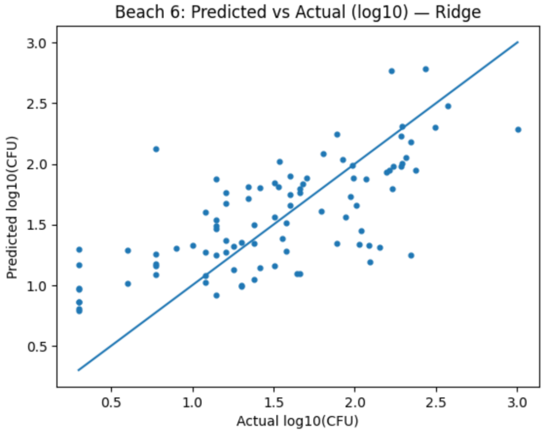

Figure 5: Predicted versus actual values for Beach 6 dataset using Ridge linear regression model.

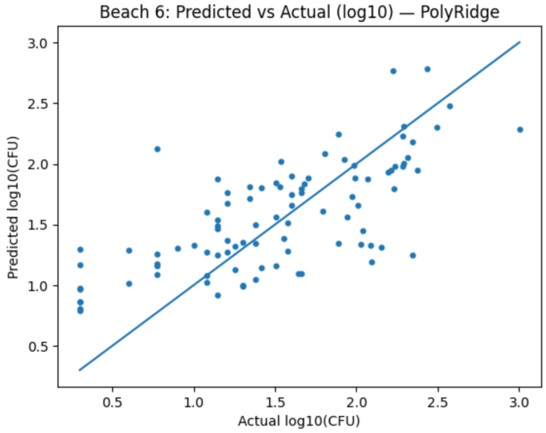

Figure 6: Predicted versus actual values for Beach 6 dataset using PolyRidge linear regression model.

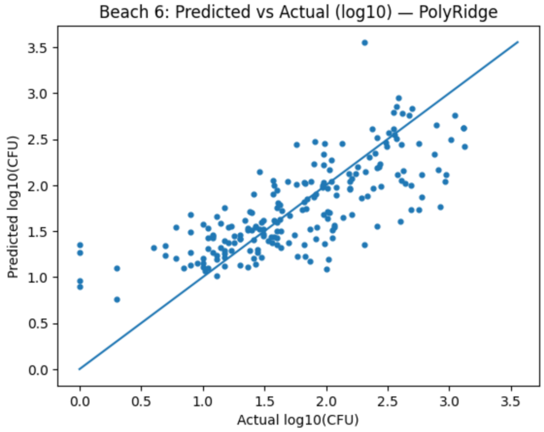

Figure 7: Predicted versus actual values for Huntington dataset using PolyRidge linear regression model.

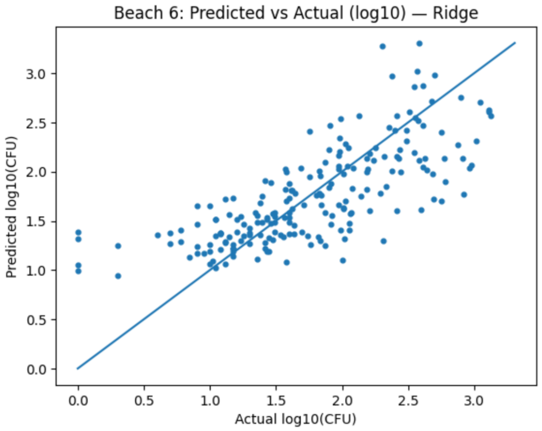

Figure 8: Predicted versus actual values for Huntington dataset using Ridge linear regression model.

experimental models into real-life applications. Using data-driven model for water quality prediction could potentially cause major public health issues if done wrong. As mentioned above, the limitation of both our models and the published one is obvious and should be used with caution when used in real life. One way to make the models more applicable is to limit the scope: for example, concentrating on predicting water quality in agricultural pond water (Matthews D. Stocker 2021). By limiting the prediction target, models could be more reliable when used for the specific scenario when data collection were standardized ideally.

6 Conclusions

Our work produces two regression models using machine learning that predict water quality in water bodies around the Great Lakes region. For the Beach 6 dataset, we found that the linear regression model using Ridge produces the most statistically successful result. For the Huntington data, the polynomial regression using Ridge generates a better prediction, as it deals with the interconnected relationships among hyperparameters. We believe that our models could be useful for progressing the water-quality machine learning model for Great Lakes region but should be used with caution when dealing with other more generalized environments. One potential future direction could be standardizing the data collection of water bodies across the Great Lakes region to produce a more generalized model with a wider scope.

7 Contributions

Both B.C. and R.W. did the early exploration and the preprocessing of the data. B.C. experimented with the first few models on the Beach 6 dataset. R.W. experimented with the Huntington dataset for the similar models. B.C. experimented cross-validation and grid search for coefficient tuning and integrated the models together for statistical analyses. R.W. was responsible for writing the textual part of the write-up. Both contributed in making figures and tables.

Both have proofread and made final editing of the document upon submission.

References

[Matthews D. Stocker 2021] Matthews D. Stocker, Yakov A. Pachepsky, R. L. H. 2021. Prediction of e. coli concentrations in agricultural pond waters: Application and comparison of machine learning algorithms. *Frontiers Artificial Intelligence* 4.

[Seok Min Hong 2024] Seok Min Hong, Billie J. Morgan, M. D. S. 2024. Using machine learning models to estimate escherichia coli concentration in an irrigation pond from water quality and drone-based rgb imagery data. *Water Research* 260.