EGU22-1833, updated on 13 Aug 2022
https://doi.org/10.5194/egusphere-egu22-1833
EGU General Assembly 2022

Estimating E. coli concentrations in irrigation pond waters with machine learning algorithms

Matthew Stocker[1], Yakov Pachepsky[2], and Robert Hill[3]
[1]United States Department of Agriculture, Agricultural Research Research, United States of America (mdstocker89@gmail.com)
[2]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee
[3]Department of Environmental Science and Technology, University of Maryland, College Park, Maryland

The microbial quality of irrigation water is an important factor in the field of food safety. Concentrations of the microbial contamination indicator, *Escherichia coli* (*E. coli*), are used to make microbial water quality determinations. However, relationships between the concentrations of *E. coli* and water quality parameters are often non-linear. Machine learning (ML) algorithms have been shown to make accurate predictions in datasets with complex relationships. The purpose of this work was to estimate *E. coli* concentrations in agricultural pond waters with several popular ML algorithms and observe the differences in model performances. Two ponds in mid-Atlantic U. S. were monitored biweekly during the irrigation seasons of 2016, 2017, and 2018. Samples were collected across the two ponds and *E. coli* concentrations were measured concurrently with 12 other water quality parameters. The resulting datasets were used to estimate *E. coli* concentrations using stochastic gradient boosting machines, random forest, support vector machines, and k-nearest neighbor algorithms. The performance of the algorithms was compared by treating performance metrics as statistics obtained by Monte-Carlo modeling of the algorithms. The results of repeated 10-fold cross-validation showed that the random forest model provided the lowest RMSE value for predicted *E. coli* concentrations in both ponds for individual years and when multi-year datasets were evaluated. However, in most cases there was no significant difference ($P > 0.05$) between RMSE of random forest and other ML models. For individual years, the normalized RMSE of the predicted *E. coli* concentrations ($\log_{10}$ CFU 100 mL^{-1}) ranged from 0.071 to 0.124 and from 0.102 to 0.155 for Ponds 1 and 2, respectively. For the 3-year datasets, these values were 0.119 and 0.132 for Ponds 1 and 2, respectively. Turbidity, dissolved organic matter content, specific conductance, chlorophyll concentration, and temperature were the most important predictors as identified by a recursive feature elimination analysis. Model predictive performance did not significantly differ when the 5 least expensive and time-consuming predictors were used as compared with the complete set of predictors. Machine learning appeared to be efficient in discerning complex relationships between *E. coli* and water quality parameters which describe the aquatic habitat.