To: Cornell Soil and Water Lab (SWL)
From: Isaiah Guenther, Biological Engineering - '24
Date: 03/17/2024
Re: Predicting pesticide contamination in groundwater

Problem Scope

Over the past three years, the SWL has been working with the New York State Department of Environmental Conservation (DEC) to test groundwater aquifers in Upstate New York for pesticide contamination. The goal is to have a better understanding of how effective current pesticide regulations are.

At the SWL, we have been analyzing the gathered data to find out what causes pesticides to leach into groundwater supplies. I propose we implement machine learning models that are taught in ORIE 4741: *Learning with Big, Messy Data* to gain a deeper understanding of the data.

Questions

- *Can we predict whether or not pesticides will leach under certain conditions? If so, can we predict the probability of leaching? Can we then predict the concentration in the groundwater?*

- *What information about the application sites and the active chemical compounds in the pesticides are most important for predicting outcomes? How can we then use theory to explain these factors?*

Data

The SWL has data on 674 pesticide tests from June of 2022 to November of 2023. Each row is a separate sample. The columns contain information about the pesticide, such as the active compound, the half-life, and the organic matter partitioning coefficient. Information about the pesticide application site is also included, such as the aquifer vulnerability, the ecoregion, and the soil drainage class. One column contains the pesticide concentration test result. Roughly half of the samples contain concentrations above zero, while the other half is undetectable.

Value

This information will help the DEC improve pesticide regulations and predict aquifer contamination. It will also help pesticide users make more informed decisions when applying. This is imperative for protecting human health, preserving our natural resources, and improving agricultural practices.

Using these machine learning algorithms will also provide another layer of verification to the on-going investigation. Different approaches are being used to analyze the data, and all of them will be compared to get the most insight possible. More tests will be conducted through 2025, so these models can be utilized on new data, as well.

—

Github - https://github.com/izguenther6/orie4741-final