Estimation of Daily Water Temperature using Random Forest

B Steele

September 15, 2023

Checkpoint Issues:

- is this a legit way to do train/val/test?!

Scientific motivation and problem statement:

Water temperature is often a reliable indicator of general water quality (cite). Active monitoring of lakes, especially those that are difficult to access by monitoring personnel, is difficult. Additionally, manual monitoring of waterbodies (by physically visiting a site) and sensor networks to monitor water temperature, are costly endeavors (cite).

By leveraging the historical manual monitoring data from Northern Water, as well as surface temperature estimates from Landsat thermal bands alongside weather data, can we adequately estimate surface water temperature using static variables (like elevation, lake area, shoreline complexity) and weather data?

In this example, I use only measured surface temperature from 7 lakes/reservoirs in the Northern Water system. 'Surface temperature' for this example is any measured temperature at >= 1m depth. I retain only the top-most value for temperature.

It's clear that there are site-level differences in temperature range and general seasonal response. These differences likely due to static variables that differentiate these lakes. That said, if I add in site-level information, the algorithm will quickly learn those key attributes and likely overfit to the data, not allowing for generalization beyond these lakes.

Training/Validation/Testing

Due to the likely influence of static variables in the algorithm my validation and testing sets will both have 'new' lakes to validate and test on.

Write up contents:

- scientific motivation and specific problem statement
- description of the data including explicit identification of the predictors and predictands
- description of any data pre-processing performed and why you did it
- training/validation/testing split
- machine learning setup and reasons for hyperparameter choices when relevant
- results (e.g. testing accuracy)
- a detailed discussion of why you don't think you have overfit
- a detailed discussion of why you think the results are better (or worse if that is the case) than a baseline approach of your choice (e.g. random chance, linear regression, climatology, etc)
- concluding thoughts including any insights gained from your efforts
- link to your github repository

4