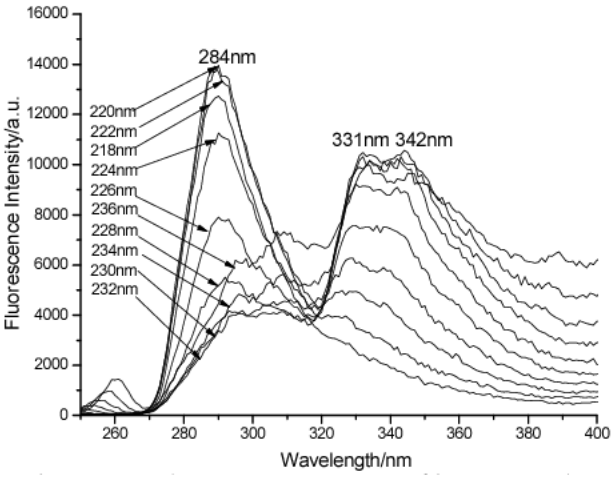
Does the water meet environmental water quality objectives set by oil and gas operations?



Fluoranthene is a contaminant whose presence in water beyond a specific concentration is harmful. Oil and gas operation require water to have a concentration of Fluoranthene <0.5 micrograms/Liter to meet water quality objectives.

To assess the quality of water, an experimental procedure is used where water is excited by different light wavelengths, and the intensity of light emitted from water as a result is recorded at different frequencies. A data set is collected which shows the radiation levels at different frequencies under different known concentrations of Fluoranthene. This constitutes a dataset that can be used to train a classifier which determines if Fluoranthene concentration is larger than or smaller than 0.5 micrograms/Liter.

Each row of the data set in the file ‘Otonabee_Fluoranthene_training’ consists of the following:

1) The first item (first column) shows the concentration of Fluoranthene in 336 water samples.
2) The rest of the row is a vectorized version of a 29x111 fluorescence spectrum of water, corresponding to 29 excitation wavelengths and 111 emission wavelengths.

This data set consists of 336 samples that can be used for training.

You are required to:

1) Convert the first column to two class labels according to the concentration ($<$ or $\geq\ 0.5$);
2) Use the remaining columns as feature vectors to train a two-class classifier according to the labels obtained in 1);
3) Use the obtained classifier to classify the samples in ‘Otonabee_Fluoranthene_test’ into the two classes. The file has 200 test samples, and is structured similar to ‘Otonabee_Fluoranthene_training’;
4) Compare with the actual class obtained by using the first column of ‘Otonabee_Fluoranthene_test’, and calculate the number of errors.
5) Discuss methods to improve your classification results.

The data has high dimensionality. Thus, you may want to group multiple emission wavelengths into one number by averaging their value into one.

Acknowledgement: *The data sets are obtained and used with permission from Dr. Nicolás Peleato (UBCO, SoE), whose research focus is on the evaluation of drinking water treatment processes, source water quality assessment, and integrated data analysis and machine learning pertaining to water quality.*

https://engineering.ok.ubc.ca/about/contact/nicolas-peleato/