Algal Bloom Forecasting in a Classification and Regression Setting

Implementing a UNet Architecture to evaluate the differences between both settings.

Author: Rodrigo Alvarez Lucendo - r.alvarezlucendo@student.tudelft.nl
Supervisors: Attila Lengyel and Robert-Jan Bruintjes
Responsible Professor: Jan van Gemert

1 INTRODUCTION

Background Information. Harmful Algal Blooms (HABs) occur when algae grow out of control and produce harmful effects on the environment, health and economy. Algae concentrations are traditionally measured via direct water sampling, a labour-intensive method limited in space and time. The research implements remote-sensing-based detection, a method that tackles these two problems and relies on predicting the estimated chlorophyll concentration as an indicator of algal bloom.

Motivation for the Research. Due to the algal's non-linear and non-stationary nature, a machine learning approach is preferred over classical models. Concretely, the UNet Architecture is used to learn the spatial features of the data. Predicting the chlorophyll concentration does not give information about the model uncertainty. Framing the regression problem as a classification problem solves this issue. Additionally, it is worth exploring the difference in performance between both settings.

Research question and sub-questions. *What are the differences between a classification and regression model for forecasting the chlorophyll-a concentration of a water reservoir?*

1. What are the differences between a classification and regression implementation of the UNet Architecture?
2. What influence does the binning strategy have?
3. How can class imbalance be mitigated using different loss functions?

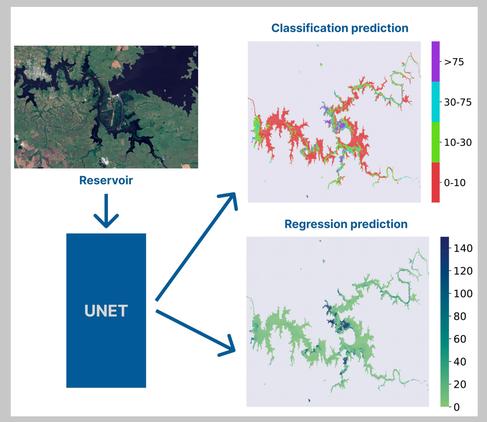

Figure 1. *The poster explores the difference between a regression and classification implementation of the UNet Architecture in the context of algal bloom forecasting.*

2 METHOD

Regression vs Classification

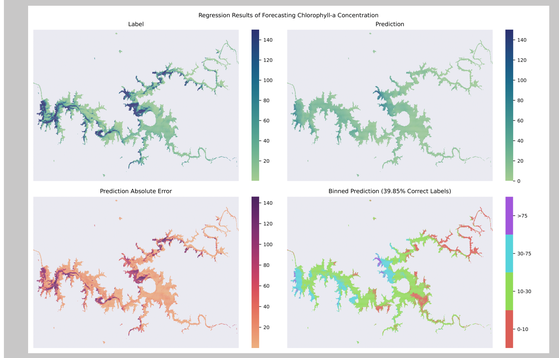

Figure 2. *Regression forecast of algal bloom. Label and prediction are the chlorophyll concentration in µg/L. Absolute prediction error and the binned regression output are shown.*

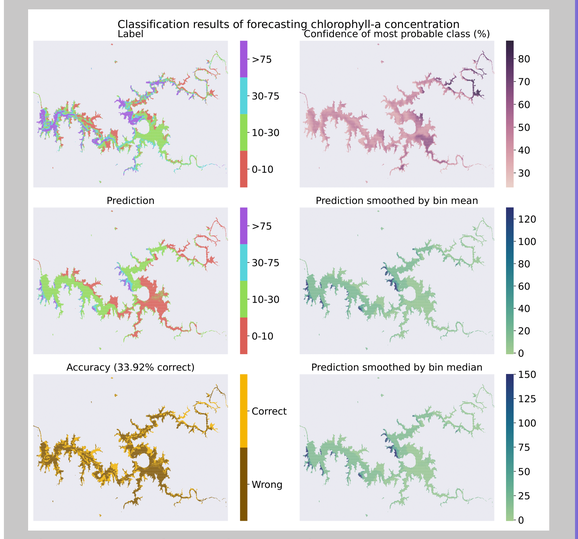

Figure 3. *Classification forecast of algal bloom. Label and prediction are discrete labels, matching labels and prediction confidence is shown. Bins are smoothed by the mean and median of each bin.*

Classification implementation. The model is trained with binned labels according to a range of values. The output is the probabilities of each class which represent the model uncertainty, and the predicted label is the class with highest probability.

Regression implementation. The model is trained normally and the output is a continuous value between 0 and 150 *µg/L*.

Binning Strategies

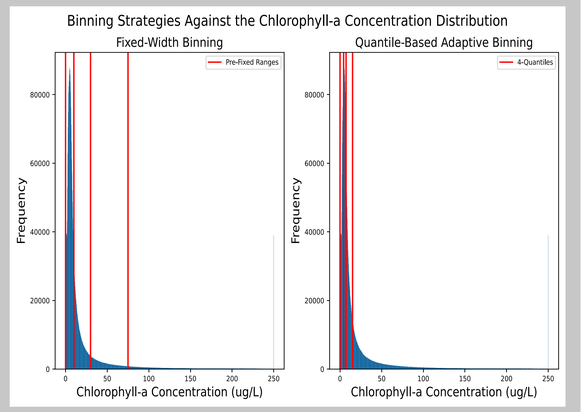

Figure 4. *Estimated chlorophyll concentration distribution and class ranges used by different binning strategies.*

Adaptive Binning creates a range of values based on the underlying data distribution, specifically Quantile-based Binning results in equal sized bins:

$$[0.0, 4.34, 7.24, 15.04, 150.0]$$

Fixed-Width Binning is based on domain knowledge given by the government of Uruguay and results in irregular bins:

$$[0, 10, 30, 75, 150]$$

In future work, a **combination** of both strategies would guarantee equal sized bins while keeping the range of values that are of interest. First, fixed-width binning will be performed and later the bigger bins would be split into more bins until all the bins contain an equal number of observations.

Loss Functions

Focal Loss. Reduces the loss more in well classified samples than in less confident misclassified samples by taking a portion of the cross entropy loss.

Dice Loss. Measures similarity between two samples by maximising the overlap or correctly classified points while minimising the union of the prediction and the ground truth.

Class-balanced Loss. Balance the loss by adding weights that are inversely proportional to the frequency of observations.

Compound Loss. Obtained by summing over different types of loss functions.

3 EXPERIMENTS

METHOD	ACCURACY (%)	MSE (μG/L)
regression	43.2 ± 0.7	1953 ± 8
classification	41.5 ± 0.9	2154 ± 35

Table 1. *Accuracy and mse validation scores for the regression and classification model using four pre-fixed bins.*

Table 1 shows regression outperforms classification using the four pre-fixed bins. A possible explanation is that se loss in regression can predict the minority classes better than cross entropy loss in classification.

BINNING	ACCURACY (%)	MSE (μG/L)
fixed-width	41.5 ± 0.9	2154 ± 35
adaptive	52.6 ± 0.6	2753 ± 10

Table 2. *Accuracy and mse validation scores for different binning strategies in classification.*

Table 2 shows adaptive binning achieves higher accuracy than fixed-width binning. However, it does not mean is better because it is not predicting the ranges of interest and may therefore not be useful in algal bloom forecasting.

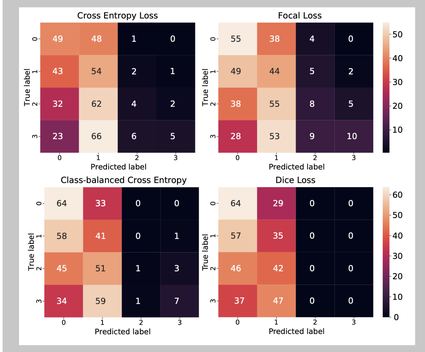

Figure 5. *Normalised confusion matrices for different loss functions in classification.*

LOSS	ACCURACY (%)
cross-entropy	41.5 ± 0.9
balanced cross-entropy	41.6 ± 0.1
focal	40.8 ± 0.4
dice	36.9 ± 0.3

Table 3. *Overall accuracy validation scores for different loss functions.*

Table 3 shows there is not significant improvement in the overall accuracy when using different loss functions.

In Figure 5 none of the loss functions learn to predict the minority classes accurately.

4 DISCUSSION

Trade-off #1. Choose between regression model that achieves higher accuracy and outputs values of higher fidelity or classification model that gives estimation of uncertainty.

Trade-off #2. Choose between adaptive binning which performs better or fixed-width binning which reveals more information about algal bloom.

Limitations. Manual tuning and UNet model not being suitable for accurate algal bloom forecasts.