1. Introduction and motivation

Understanding the Phytoplankton growth is essential as it is the base of the whole biological food chain in the ocean, as Australia is a country surrounded by ocean. There are many ocean gliders that collect different properties of the water, such as temperature, pressure, latitude, etc. The aim for this project is to find the related features that affect the chlorophyll-a, which can be found in phytoplankton, and use machine learning and statistical methods to predict the growth of concentration of phytoplankton. We will both use Rstudio and Python to pre-process and visualize data, fit the data to various forms of regression models, draw comparisons between predicted values with actual results. Ultimately, select the best fitted model and predict the biological growth.

1. Literature Review

The principle problem of our project revolves around the analysis of the relationships between a selection of variables and the concentration of chlorophyll-a (hereafter referred to as CPHL), this is based off the assumption that CPHL is a valid proxy for the presence of algal biomass (phytoplankton). Such a presence of biomass is a focal point of the project and has implications upon fields of research outside our scope.

To begin with we can examine the relationship between CPHL and algal biovolume. In the literature, CPHL has been widely used as a proxy for algal biovolume, in a 1981 study; “Relationship between Chlorophyll-a Concentration and Phytoplankton Biomass in Several Reservoirs in Czechoslovakia”[1] researchers obtained samples from 5 different water reservoirs at 3-week intervals. Calculations of concentrations and phytoplankton biomass values were used in linear regressions to analyse the co-variation of the two measures finding significant positive correlation with values in the ranges of 0.75-0.9 for most reservoirs, providing strong evidence for the existence of a relationship. Similarly, a study on the relationship from 2007; “Does chlorophyll a provide the best index of phytoplankton biomass for primary productivity studies?”[2] compared the chlorophyll-a proxy to other five other proxies, ultimately finding that CPHL provided an equal or more accurate estimate than proposed proxies, on the basis of correlation coefficient, root mean square error and mean absolute percent error.

Examining some of the relationships between our own data and CPHL we can review some of the literature exploring the variation in surface chlorophyll-a concentration. Two case studies: “Surface chlorophyll concentrations in relation to the Antarctic Polar Front”[3] and “Seasonal variation of chlorophyll-a concentration … in the subtropical East China Sea”[4] use satellite observations of sea surface temperatures (SST) and ocean colour data to examine the fluctuation of CPHL. Both cite the light limitations of certain seasonal time periods as a major factor in CPHL variation. Due to solar declination during the winter months surface radiation limits the growth of algal blooms causing a decrease in CPHL. A separate study on the East China Sea: “Evaluating the impact of sea surface temperature (SST) on spatial distribution of chlorophyll-a concentration in the East China Sea”[5] used the same SST and ocean colour observations to find significant positive correlation between surface temperatures and CPHL on the north side of the East China Sea, but found negative correlations in the south, citing low nutrient density in the area as a possible cause.

1. Methods, software and Data Description

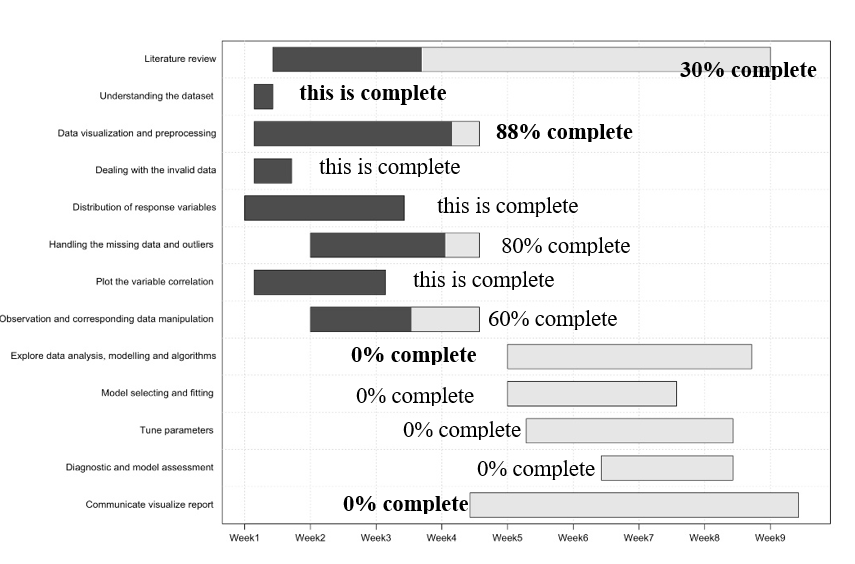
The given raw dataset is in the CSV format with 1.16 GB storage. There are 3123117 observations with 58 variables, including the numerical response CPHL. We aim to use the remaining numerical variables and their corresponding classifications to predict the value of CPHL, which is the measurement of phytoplankton concentration.

It can tell from the dataset that there are seven distinct gliders recorded data from seven different periods between 2013 and 2015, while the gliders were placed at various depths, latitudes, and longitudes. Nevertheless, the shortage of raw data is noticeable. A large number of data at around 19% is missing that spreads unevenly across the variables. Take one variable, VCUR, which is the value of seawater velocity at northward as an example, a significant amount at 97% of the data is missing. It is crucial to balance the importance of missing data and the information it has contained during data preprocessing. Another difficulty of the dataset is the interaction between different variables. The giant size of variables will quickly impact the accuracy of correlation detection so that it is vital to exclude the noise from disturbing features in data analysis.

We intend to apply Python and R on the project, while Python will mainly concentrate on missing data filling, and R will focus more on modeling and diagnostic. Our first step is to visualize the data and clear out the invalid value range. It is helpful to plot the correlation matrix with the library("corrplot"),  scatterplot with a smooth spline and histogram to evaluate the potential relationship between response and predictors, and the percentage of the data quality (i.e., good or bad data). Different methods will be tested via python to deal with missing values, including mean, mode, KNN, random forest filling ways with varieties of sklearn libraries. After comparing the replacement accuracy, the best performance approach can be taken to acquire the completable without missing value. Then it goes to the other essential process that is the data modeling and algorithms. Due to the positively distributed numerical response CPHL, it is highly likely that the topic is a regression model issue. In that case, we would like to test different models' performance, including xgboost, random forest, and elastic net by python with higher efficiency than R but mainly concentrates on regression models containing logistic, lasso, and support vector regression in R due to the precision. Residual vs. fitted value plot and QQ-plot will help us to diagnose the models. Besides, we also have the hyperparameter tuning plans to improve the accuracy after comparing measurements such as MSE, R-squared, and AUC-ROC diagrams and Anova tables.

The above processes on the modeling build will be assessed in R and Python to take advantage of both, and a better result with higher performance one can be selected.

1. Activities and Schedule



**Timetable for the activities and the percentage of task completion by week4**

References

[1] Desortová, B., 1981. Relationship between Chlorophyll-α Concentration and Phytoplankton Biomass in Several Reservoirs in Czechoslovakia. *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, 66(2), pp.153-169.

[2] Huot, Y., Babin, M., Bruyant, F., Grob, C., Twardowski, M. and Claustre, H., 2007. Does chlorophyll-α; provide the best index of phytoplankton biomass for primary productivity studies?. *Biogeosciences Discussions*, 4(2), pp.707-745.

[3] Moore, J. and Abbott, M., 2002. Surface chlorophyll concentrations in relation to the Antarctic Polar Front: seasonal and spatial patterns from satellite observations. *Journal of Marine Systems*, 37(1-3), pp.69-86.

[4] Gong, G., Wen, Y., Wang, B. and Liu, G., 2003. Seasonal variation of chlorophyll a concentration, primary production and environmental conditions in the subtropical East China Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, 50(6-7), pp.1219-1236.

[5] Ji, C., Zhang, Y., Cheng, Q., Tsou, J., Jiang, T. and Liang, X., 2018. Evaluating the impact of sea surface temperature (SST) on spatial distribution of chlorophyll-a concentration in the East China Sea. *International Journal of Applied Earth Observation and Geoinformation*, 68, pp.252-261.