
UNSW SYDNEY

# CHOICE OF A PROBLEM BY TEAM 12

# A DATA SCIENCE APPROACH TO FORECAST PHYTOPLANKTON CONCENTRATION

Xinyu Xu (z5175081), Yuewen Mao (z5210649), Zilin Li (z5158442), Cheng Qian (z5158272), Henry Jiang (z5205963).

School of Mathematics and Statistics
UNSW Sydney

October 2020

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF
THE CAPSTONE COURSE DATA3001

## Abstract

With roughly 85% of Australia's population living on its coast the ocean is an important and ever-present factor in our lives. Phytoplankton, being the core building block of many marine food-webs, has significant influence over the health of marine biospheres and indirectly affects our lives and wellbeing. As such an exploration of phytoplankton and how it is affected by environmental factors is an important course of action to better understand marine ecology and the focus of this project. In order to observe the presence of phytoplankton, the agreed upon standard is the observation of chlorophyll- α concentration, an important molecule used in the photosynthesis process. Environmental factors such as temperature and light intensity are included in our dataset and we can use these to produce and compare models of chlorophyll-α concentration using R. However, our dataset poses some difficulty in the pre-processing phase, it contains large chunks of missing data which we intend to predict or remove depending on the circumstances through the use of Python. We are expecting that a regression model will provide the best fit and are looking to examine different modelling techniques. In order to accomplish this project efficiently and effectively, we have split the workload into sections, with focus on understanding the data, reviewing literature, visualizing relationships and the data, exploring and modelling data and communicating the results in a final report.

# Contents

## 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. Many gliders collect different properties of the sea, such as temperature, pressure, etc. This project aims 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 use Rstudio and Python to pre-process and visualize data, fit the data to various forms of models, draw comparisons between predicted values with actual results. Ultimately, select the best-fitted model and predict the growth.

## 2. Brief Literature Review

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

To begin, we can examine the relationship between CPHL and algal biovolume. In relevant literature, CPHL has been widely used as a proxy for algal biovolume. A study on the validity of the relationship from 2007; "Does chlorophyll a provide the best index of phytoplankton biomass...?"[1] compared the chlorophyll-a proxy to other five others, ultimately finding that CPHL provided an equal or more accurate estimate on the basis of correlation coefficient, root mean square error and mean absolute percent error.

Concerning relevant environmental variables both "Surface chlorophyll concentrations in ... the Antarctic Polar Front"[2] and "Seasonal variation of chlorophyll-a concentration … in the East China Sea (ECS)"[3] used satellite observations of sea surface temperatures (SST) and ocean colour to examine fluctuations of CPHL. Both cite the light limitations (a result of solar declination) of certain seasonal periods as a major factor in CPHL variation. A separate study on the ECS: "Evaluating the impact of SST on spatial distribution of chlorophyll-a concentration in the East China Sea"[4] used the same observations to find significant positive correlation between SST and CPHL in the northern ECS, but found negative correlation in the south, citing low nutrient density in the area as a possible cause.

## 3. 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, 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 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.

## 4. Activities and Schedule

The following are the main project activities, including five main components:

**-Understanding the dataset (wk1)**

*Acquisition of the dataset and understanding the representation of different variables.*

**-Literature review (wk2-wk8)**

*Searching the literature is conducted during the whole project.*

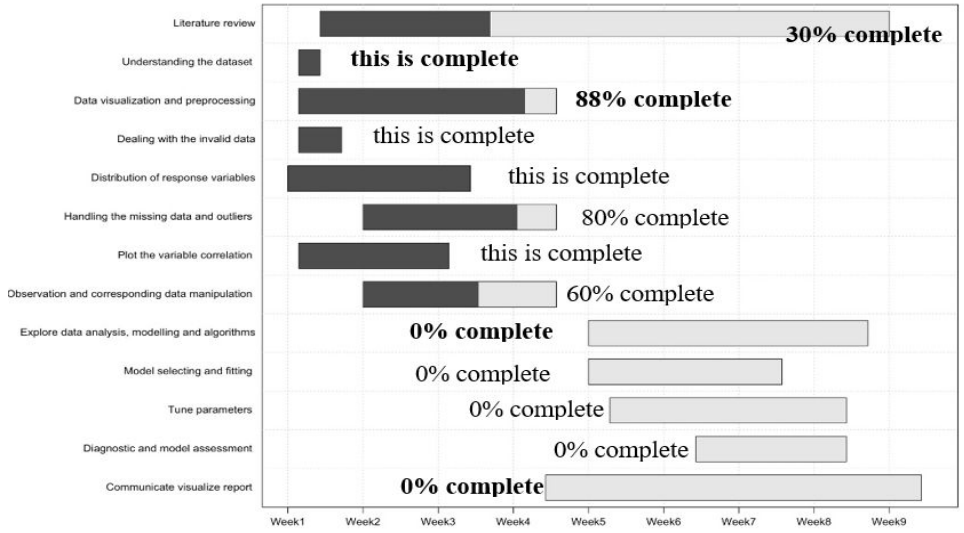
**-Data visualization and preprocessing (wk2-wk5)**

*This is the most time-consumption part of our project. We aim to test the accuracy of different imputations methods to fill missing values. It includes the following steps: Dealing with invalid data, Distribution of response variables, Handling missing data and outliers, Plot the variable correlation diagram, Observation, and corresponding data manipulation.*

**-Explore data analysis, modelling and algorithms (wk6-wk8)**

*Selecting the related modelling and tuning parameters to increase accuracy, choosing the fittest models after the assessment.Model selecting and fitting, Tune parameters and Diagnostic and model assessment are included in this part.*

**-Communicate visualize report (wk5-wk9)**




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

# Reference

[1] 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.

[2] 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.

[3] 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.

[4] 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.