Prudhomme Pierre - Berton Thomas
Groupe 19
November 15th 2022

MRR Project : Ocean

Introduction

The chlorophylle-A is a pigment which is present in every plant, it allows the plants to receive light to start photosynthesis. In Oceanography, the chlorophylle-A quantity in the water tells us about the quantity of phytoplankton in the water.

Presentation of the dataset

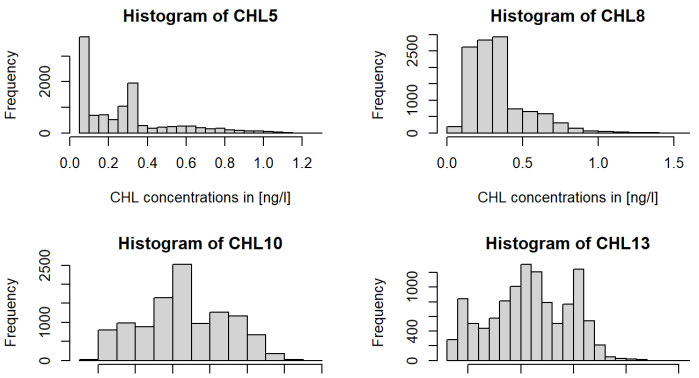
The dataset has 11169 observations of 44 variables. Each variable represents a measure of a slice into the water. The more we go on with the columns the more we go down in the depth of the ocean. We need to focus on a zone called BATS and work on those specified coordinates. Those data have been collected from 1992 to 2008, and they are divided by 73 sequences of 5 days. The dataset contains the average vertical Chlorophyll-a distribution at 17 depth levels (from 5 to 217 m) and the temperature at these depth levels. In addition to the vertical profiles, we also have surface variable values such as SSH, CC, WS, SR. These values constitute the components of the model surface vectors.

Description of the variables

This dataset contains :

- **SSH :** Sea-Surface Height
- **CC :** Cloud Cover
- **WS :** wind speed
- **SR :** Downwelling Shortwave Radiation
- **Therm k :** Vertical Thermic Profile with k from 1 to 18 (18 different depth from 5 to 217 m)
- **CHLk :** Chlorophyll-a Concentration with k from 1 to 18 (18 different depth from 5 to 217 m)
- **Year :** The year where the measure was taken.
- **Latitude :** Latitude where the measure was taken.
- **Longitude :** Longitude where the measure was taken.
- **5days :** The number of the 5days sequence.
-

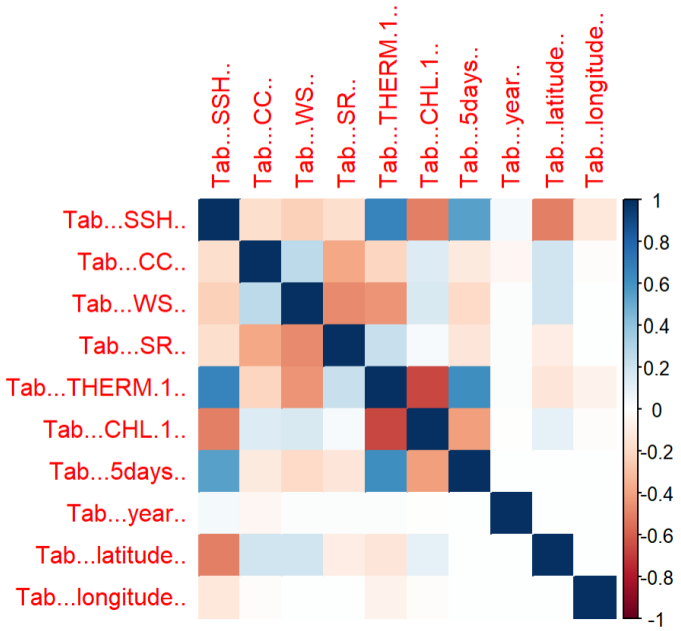
Due to n = 11169 observations and p = 44 variables we can assume that n >> p.



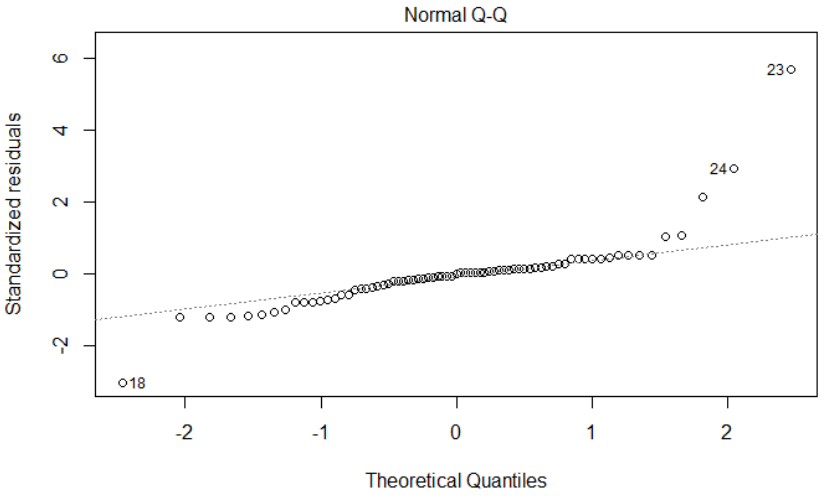
We choose to print some of the Chlorophyll-A concentration to build a First idea of what they will look like.
As you can see for the first slice, the concentration is very badly distributed but this will be more scattered as soon as we go down.

Correlation Matrix :

As we can see here we have the Correlation matrix built by the Library ggplot.
And we choose not to take all of the Variables, we only choose the “first” Variables the measure on top of the Ocean, and some others. We can see that there are several linearities. Especially between the different measures of chlorophylle-A. Therefore, we can reduce the number of factors in our relations.

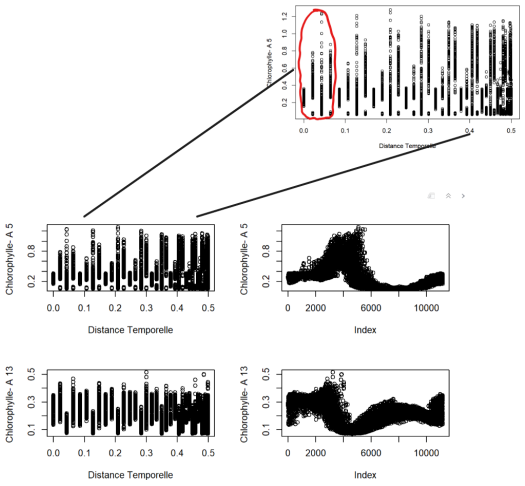


First linear regression :



We have done a quick regression to try out our model.
We have modelise the chlorophyll measure of one month with all other first variables.
And we can see, thanks to this QQ plot that our modelised variable is following a normal distribution.

Periodicity :



Finally, we wanted to study the periodicity of the variables. As we can see on the graph where we have the Chlorophylle-A 5 distribution over the temporal distance, we can see that the model is repeating. Therefore, the Chlorophylle-A distribution is periodique. This periodicity was predictable because the Chlorophylle-A distribution is naturally periodical because of the season.

$$Temporal\ Distance\ = \sqrt{sin(2\pi x/73)^2\ *\ cos(2\pi x\ /\ 73\ )^2}$$