Phytoplankton time series modelling

Robin

October 20, 2020

The aim of the approach is to characterise the relationships existing between the phytoplankton functional groups (PFTs) and their environment with a focus on extreme environmental events. Extreme events can be understood as extreme values/brutal changes in the values of some environmental covariates such as temperature or wind. More precisely the emphasis will be laid on the length and magnitude of the reactions of the organisms to their environment.
The interest of such a study is twofold:

- Better characterise the frequency of the data needed to study phytoplanktonic communities. For instance, if the reaction delay of the PFTs appears to be very short then this work would advocate for more high-frequency series acquisitions.
- If the relationship existing between the phytoplankton and the covariates is better understood then in a second time one could estimate the impact of each covariate on the quantity of carbon synthesised by the phytoplankton cells. It can be wondered for instance whether extreme events account for a significant share of the POC export even if they occur by definition rather rarely.

The explanatory power of the chosen model is here more important than its predictive power. As such, the statistical methods used will have to remain interpretable.

Commenté [GG1]: To define : in terms of what ? Abundances? Cell size? Cell cycle?

Commenté [GG2]: Don't we already know that ? I mean if for instance one cell divide once a day, then making a measure twice a day would be enough (Nyquist).

Wikipedia : La **fréquence de Nyquist**, du nom de l'ingénieur électronicien Harry Nyquist, est la fréquence maximale que doit contenir un signal pour permettre sa description non ambiguë par un échantillonnage à intervalles réguliers. Elle est aussi connue sous le nom de fréquence limite de repliement. Elle est égale à la moitié de la fréquence d'échantillonnage.

Commenté [GG3]: Of which variables ? The phyto ? The environment?

Commenté [GG4]: What do you mean ? The C export from the surface to the deep ocean and sediment (C sequestration for long term)?

The fate of C is very complex and rather than talking about the export I would rather focus on the "potential" C fixed by the phytoplankton biomass. Then the fate of this biomass is a different story you cannot address with your data as there is no information about predators, viral lyses, sediment traps.

1 Data

We have two types of data: Low (twice a month) and high frequency (6 times per day) data which both include the PFT time series and ~~the~~ various environmental covariates time series (listed below).

1.1 SOMLIT data

The SOMLIT series has started in 1997 with one marine station and nowadays counts 10 stations. The SOMLIT Network is a National Obersvatory Service (https://www.somlit.fr/). The measurements are made every two weeks which makes this series a low frequency time series. We will ~~keep here~~focus only on the station ~~existing~~ in Marseille (the ~~Frioul~~ SOLEMIO station) that collects flow cytometry data from 2009 to nowadays. Other variables have been collected for more than 20 years.

1.1.1 The dependent variables

Five PFTs are tracked: Cryptophytes, ~~synechoccocus~~Synechoccocus, ~~prochlorocchocus~~Prochlorocchocus, picoeucaryotes and nanoeucaryotes.

1.1.2 The covariates

The covariates recorded in the SOMLIT dataset are: Temperature, Salinity, oxygen (O), pH, Ammonium (NH4), Nitrates (NO2-), Nitrites (NO3-), Phosphates (PO4), Silicates (SIOH4), Particular Organic Carbon (POC), Particular Organic Nitrogen (PON), Material in Suspension, A-chlorophyll, Delta N15, Delta N13.

1.2 SSLAMM station data

The time frequency of the SSLAMM series is of an acquisition every two hours starting from mid-September 2019 until now. There are four ”holes” in the data (each ranging from three days to one month and a half).

1.2.1 The dependent variables

For the moment we consider six groups to track (the nomenclature might evolve). These are the same five PFTs as in the SOMLIT dataset: picoeucaryotes, ~~synechoccocus~~Synechoccocus, ~~prochlorocchocus~~Prochlorocchocus, nanoeucaryotes and cryptophytes. In addition to these five PFTs the microphytoplankton group is here tracked.

1.2.2 The covariates

The temperature was recorded *in situ* since mid-September 2019 on a ~~one~~ hour basis. In addition, silicate, nitrite, phosphate, nitrate and ammonium concentrations have also been measured *in situ* every four days on average. Besides, variables outputted by two oceanographic models, WRF and MARS-3D, such as the Air temperature, wind direction and strength, sun flow (W/m^2), the salinity and the horizontal water currents with a one hour frequency can be used. Some variables are available both as *in situ* data and as models output: we could use them as an harmonisation basis.
To summarize, for the SSLAMM data the available variables are:

- Meteorological: Air temperature, wind direction and strength, sun flow (W/m^2).
- Oceanographic: Water temperature, salinity, turbulence, horizontal water currents, pH (pH to check).
- Biological: NH4, Nod, Nox, Nop, C, phosphorus, $NO2-$, SIOH4,PO_4^{3-}, $NO3-$.

Commenté [GG5]: I would place these meteorological data apart as they can be used for both SSLAMM and SOMLIT.

Model are very sensitive to the conditions to the limit. And the SSLAMM marine station is nearby the shore, so close to the limit of the model for MARS-3D. How do you deal with that? Do you consider an average situation for the Bay that you apply to both SOMLIT and SSLAMM data?

Commenté [GG6]: What is the difference btween Phosphorus and PO4??? And NO3 and Nod?

1.3 Summary tables

The following two tables give the variables in common/not in common between the two datasets.

	SSLAMM	not in SSLAMM
SOMLIT	Cryptophyte, synechoccocus, prochlorocchocus, picoeucaryote, nanoeucaryote.	∅
not in SOMLIT	Microphytoplankton.	**X**

Table 1: PFTs tracked comparison between SOMLIT and SSLAMM

Commenté [GG7]: But in the Phytobs database.

	SSLAMM	not in SSLAMM
SOMLIT	Water temperature, Salinity, Ammonium (NH4), Carbon/POC, Nitrates ($NO2-$), Nitrites ($NO3-$), Phosphates ($PO4-$), Silicates ($SIOH4-$), pH.	oxygen (O), Particular Organic Nitrogen (PON), Material in Suspension, A-chlorophyll, Delta N15, Delta N13.

not in SOMLIT	Air temperature, wind direction and strength, sun flow, turbulence, horizontal water currents, Nod, Nox, Nop, phosphorus.	**X**

Table 2: Covariates tracked comparison between SOMLIT and SSLAMM

2 Methods

2.1 General outline

The major steps of the project:

- **Shift analysis** on SOMLIT data. Did the covariates/dependent variables have experienced permanent regime changes since ten years ?
- **Seasonality and long trend patterns identification** on SOMLIT data. Once these patterns extracted, we could use them as additional variables in the SSLAMM data.
- **Extreme events identification** on SSLAMM data.
- **Variable selection**: The same methodology as in [4] could be used: k-fold Cross-Validation, Feature Selection, Cumsum, Gradient Analysis. Instead of gradient analysis we could also use Sobol indices. In order to check if the models correctly identify the relevant variables we could include a purely random time series as a variable and check if it is selected.
- **Identify the reaction times of the organisms:** Check the autocorrelations (how each PFT is correlated with its past values) and cross-correlations (how each PFT is correlated with the other PFTs) of the dependent series. It could be informative of the number of time lags of dependent variables to include in the model.
- **Assess the model predictive power** with cross-validation methods (dividing the series in several periods, train the model over a period and test it over the following one). The metric could be the regular Normalised Root Mean Square Error (RMSE). Yet, we could give more weights to errors occurring during extreme events to force the models to better capture them.
- **Qualititive comparison of the results**: Our results will be compared with other studies that deals with the influence of the environment on the PFTs. The major difficulty of these comparisons is that the authors often use different PFT nomenclatures than ours. Some references that could be use to compare the results with are given in the bibliography ([1], [2], [3], [5], [7]).

2.2 Models

The following models and packages could be used:

Model	Package
Distributed lag selection model (explanatory benchmark)	dlnm (R)
VARIMA + lag penalisation (explanatory benchmark)	statsmodels (Python), MTS (R)
Logistic regression for time series (if we study the relative abundance of each group).	Scikit-learn (Python)
Recurrent Neural Networks : LSTM, GRU (predictive benchmark)	Keras + tensorflow (Python)
Genetic Programming	Deap (Python)

Automatic change detection	Ruptures (Python)

3 Open issues

- **Good ~~average~~ performance on average situation vs good performance on extreme events**. There is a trade-off between having a good explanatory model on average and a model that describes well events that occur very rarely in the series (as it is the case for extreme events). Several options are possible: modelling the average and extreme events with the same model, making one model (or component) to describe the trend and another one to model the deviations from the trend or finally using two distinct models, one for ”normal” events and another one for ”extreme events”.

- **How to combine the two data sources ?** We have two types of series: Low and high frequency series. Low frequency data can be used to identify the main trend/seasonal effects existing over the long run so that high frequency patterns could then be more precisely identified. It is also possible to fit the same models over the two series (high and low frequency). If one observes different results over the two series then the analysis might be time-unit dependent/non-fractal, which is an issue in itself. The simplest way could be to extract seasonal components from SOMLIT data and add them as covariates in the SSLAMM regression problem.

- **Relative abundance vs absolute abundance** Do we track the relative abundance of each PFT among all PFTs or the absolute number of cells of each PFT in a given water volume ? The share of each PFT among all PFTs might better account for the fact that the PFTs are in competition in the environment.

Commenté [GG8]: It is some kind of normalization in order to deal with the shape of the fluctuation rather than its magnitude.

- **Non-stationarity/ non-ergodicity**: The relationship between covariates and dependent variables might change through time. This could be hard to handle for statistical model, how is it for GP ?

- **Endogeneity**: The link between the PFTs and the environmental variables is not unidirectional. In other words, the environment influences the phytoplankton but the phytoplankton also acts back over some of the variables of the environment (for instance over the Ammonium concentration). It is then difficult to identify a causal link between PFTs and the environment, which is a problem for standard statistical models. How does it impact the GP approach ?

- **Deal with the significant part of missing values in the SOMLIT data**.

- **Univariate vs Multivariate optimization.** Should we minimize the distance between the predicted and actual series for all six series in the same minimisation program or make six different minimisation programs ? Computing the correlation matrix between the series might give a small clue about this question.

- **Bloat control**. How to effectively control bloat for GP. Should we include the minimisation of the size of the solution as an explicit objective in the program ?

- **Change detection criteria** What criteria should we use to detect extreme events ? We could look for the x% most extreme values/variations of each covariate. We could also try to automatically distinguish breaks of trend/variance in the data [6]. The automatic break detection seems the most appropriate. Indeed, by taking the x% highest values/variations we might capture some outliers in the covariate series.

4 Bibliography

References

[1] Simone J Cardoso et al. “Environmental factors driving phytoplankton taxonomic and functional diversity in Amazonian floodplain lakes”. In: *Hydrobiologia* 802.1 (2017), pp. 115–130.

[2] Luciane Oliveira Crossetti et al. “Is phytoplankton functional classification a suitable tool to investigate spatial heterogeneity in a subtropical shallow lake?” In: *Limnologica* 43.3 (2013), pp. 157–163.

[3] Chengxue Ma et al. “Spatial and temporal variation of phytoplankton functional groups in extremely alkaline Dali Nur Lake, North China”. In: *Journal of Freshwater Ecology* 34.1 (2019), pp. 91–105.

[4] Danny J Papworth, Simone Marini, and Alessandra Conversi. “A novel, unbiased analysis approach for investigating population dynamics: A case study on Calanus finmarchicus and its decline in the North Sea”. In: *PloS one* 11.7 (2016), e0158230.

[5] Chang Tian et al. “Phytoplankton Functional Groups Variation and Influencing Factors in a Shallow Temperate Lake: Tian et al.” In: *Water Environment Research* 90.6 (2018), pp. 510–519.

[6] Charles Truong, Laurent Oudre, and Nicolas Vayatis. “Selective review of offline change point detection methods”. In: *Signal Processing* 167 (2020), p. 107299.

[7] Barbara Furrigo Zanco et al. “Phytoplankton functional groups indicators of environmental conditions in floodplain rivers and lakes of the Paran´a Basin”. In: *Acta Limnologica Brasiliensia* 29 (2017).