OGS
Istituto Nazionale
di Oceanografia
e di Geofisica
Sperimentale

MULTIVARIATE RELATIONSHIP IN BIG DATA COLLECTION OF OCEAN OBSERVING SYSTEM

Speaker : Pietropolli Gloria, PhD student at University of Trieste

Mail : gloria.pietropolli@phd.units.it

OUTLINE

1. **INTRODUCTION** : Why we need to observe Ocean's variable?
2. **DATASET** used for the model: the EMODnet
3. **NEURAL NETWORK ARCHITECTURE** for create the model
4. A new technique to **IMPROVE THE QUALITY OF THE DATASET**
5. **RESULTS**
6. Prediction's Confidence Interval
7. **CONCLUSION**

INTRODUCTION

Observing the ocean gives us the essential informations necessary for our sustainable development, safety, wellbeing, and prosperity.

Marine ecosystem health is impacted by human activity, in fact during the last decades, the ocean is increasingly affected by global changes

More specifically,
the Mediterranean Sea is considered an ocean in miniature, as it is distinguished by peculiar biogeochemical characteristics

One of the principal consequences is that the Mediterranean is more prone to adsorb and store anthropogenic carbon

HOW TO MEASURE MARINE VARIABLE?

Historically, the measurements of the marine variable were performed by **specific cruises** that sailed around the ocean gathering water samples and subsequently analyzing them in laboratory.

Limits:

- High cost of this shipping
- Small amount of data that this boat can gather with respect to the effort made to collect them
- Huge gap between the number of spring's or summer's measurements and the autumn's or winter's ones

FLOAT

A ***profiling float*** is a two meters long robotic device that collects marine variable data diving in the ocean varying its depth changing the buoyancy, then it rises to the surface and send the data to the stations via satellite and repeats the process.

Strength :

- Do not need human operating or any ship involvement

Limits :

- The measurements result less precise than the boat's ones, also because some sensors may decade after some years rendering the relative acquired data inaccurate or indeed incorrect.

MACHINE LEARNING CAN FILL PRACTICAL GAPS

Float sensor measures **temperature**, **salinity** and **pressure**. For each other variable, such as nitrate, phosphate, silicate and so on, a specific sensor must be inserted and the cost of building the float instruments raises drastically.

Hence, are available more information about variables such as temperature, salinity and pressure, whereas **other sea's parameters** are measured **less frequently** in fewer Mediterranean areas.

The idea is to use **machine learning techniques** to **fill these practical gaps**, exploiting the big amount of marine data that has been collected during the last decades

- For the **prediction** of some of the ocean's variables that would be too expensive to achieve for a float instrument
- For **checking** if the prediction results reliable in a situation where the sensor is present on the float.

The EMOdnet Dataset

The **EMODnet** (*European Marine Observation and Data Network*) is the long-term marine data initiative started by DG MARE in 2009, created with the aim to render marine data easily accessible, interoperable and free on restrictions on use.

The EMODnet Chemistry portal describes marine data until 2018, acquired from research cruises.

The subsetof EMODnet containing information about the **Mediterranean Sea** consistsof a collection of 101.526 vertical profiles, originated from 74 organization distributed among 18 countries.

The collected data ranges in longitude from −5,92 to 36,19 and in latitude from 31,19 to 45,77, guaranteeing a good coverage of the whole Mediterranean area.

The parameters included ineach measurement include date and geolocation of the sampling, temperature, salinity and oxygen density.

Moreover, when available, these samples can contain macronutrients such as Nitrates, Phosphates, Silicates, carbonate system variables such as the total Alkalinity, also Chlorophyll-a, Ammonium and Phosphate.

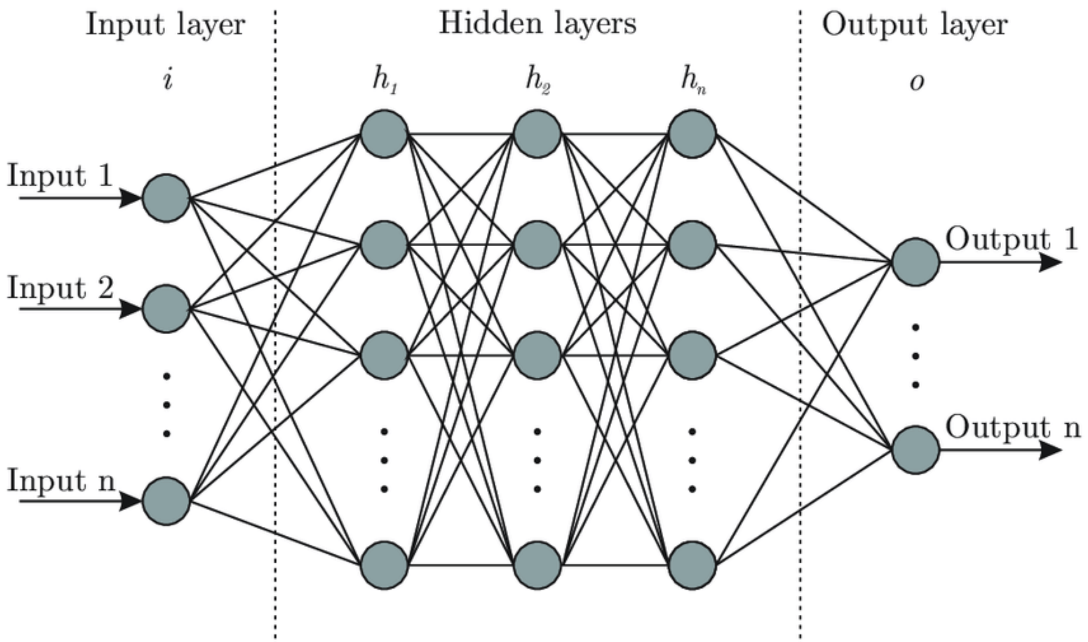
OGS
Istituto Nazionale
di Oceanografia
e di Geofisica
Sperimentale

The Multilayer Perceptron

Multilayer Perceptron (MLP) is a *Feed-Forward Artificial Neural Network* composed by a **fixed number of layers of nodes** connected between them as in a direct graph between the input and the output layer.

For the training, the **backpropagation algorithm** is utilized.

The weights and biases of the MLPs are updated during every epoch.



Implementation Details

- **Ten optimal topologies** Multilayer Perceptrons are introduced. The final output of our model is the **mean** of the ten results obtained.
- For training and validate the network are utilized measurements from **EMODnet dataset**.
- **Inputs**: date and geolocation of the measurement, temperature, salinity and oxygen density.
- **Outputs**: nitrates, phosphates, silicates, total alkalinity, Chlorophyll-a, Ammonium and Phosphorus.
- The dataset is split in a **training set**, used to optimize the weights and biases, and a **validation set** used to test the performance of the proposed model.
- the input and the output data are **normalized** in order to make training faster and to reduce the chances of getting stuck in local optima.
- The **metrics** utilized to evaluate the performance of the models are the MAE (Mean Absolute Error) and the RMSE (Root Mean Square Error).

Improvement of the quality of the Dataset

Applying the resulting model to the validation set, we observed the presence of some **outlier output**, i.e. data over which our network's prediction fails drastically.

The idea is to **clean** the dataset from these incorrect observations working directly on the neural network outputs: if our model computes a prediction that is completely wrong we will provide to its **removal** by deleting directly all the data measured by the boat responsible of this outliers both from training and validation set.

Improvement of the quality of the Dataset

Process of elimination of problematic stations:

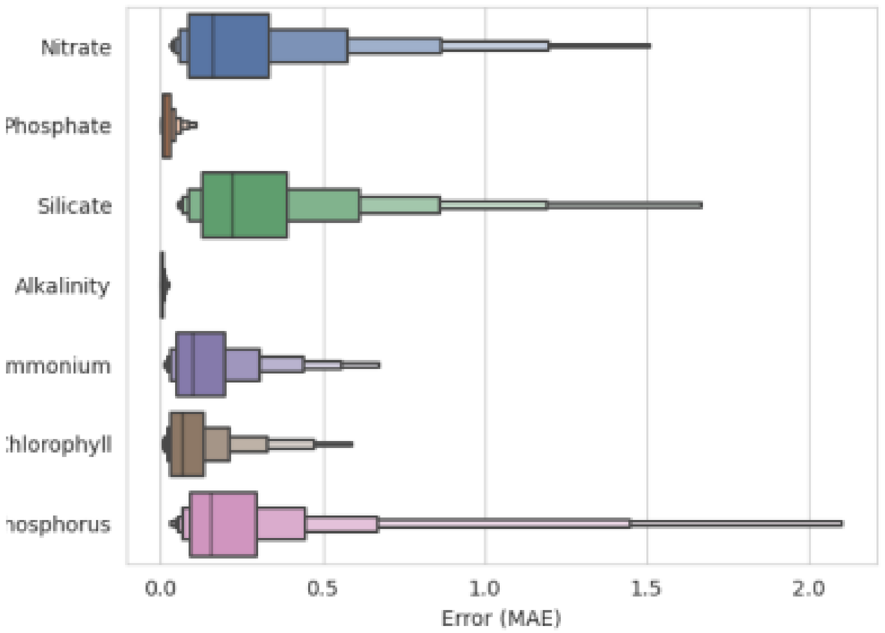
1. The models are **trained** using the **whole dataset**
2. The boat that contained an elevate quantity of samples with extremely high error in the prediction are **selected** and **removed**
3. The models are **trained** over the **new dataset**
4. The process is **repeated**

During the whole process the goal is to find an **equilibrium** between:

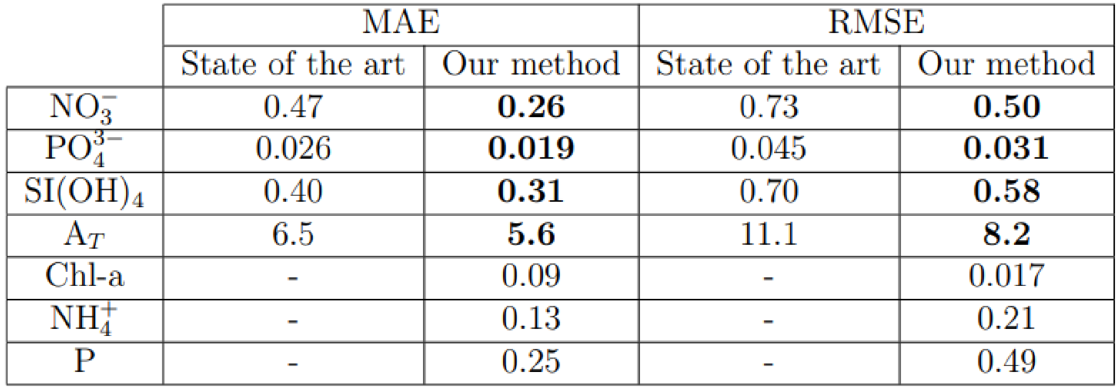
- Necessity of delete problematic data
- Necessity to avoid the removal of boats that leads important information for the prediction.

OGS Istituto Nazionale di Oceanografia e di Geofisica Sperimentale

Results



Boxplots of the testing fitness.



	MAE		RMSE	
	State of the art	Our method	State of the art	Our method
NO_3^-	0.47	**0.26**	0.73	**0.50**
PO_4^{3-}	0.026	**0.019**	0.045	**0.031**
$SI(OH)_4$	0.40	**0.31**	0.70	**0.58**
A_T	6.5	**5.6**	11.1	**8.2**
Chl-a	-	0.09	-	0.017
NH_4^+	-	0.13	-	0.21
P	-	0.25	-	0.49

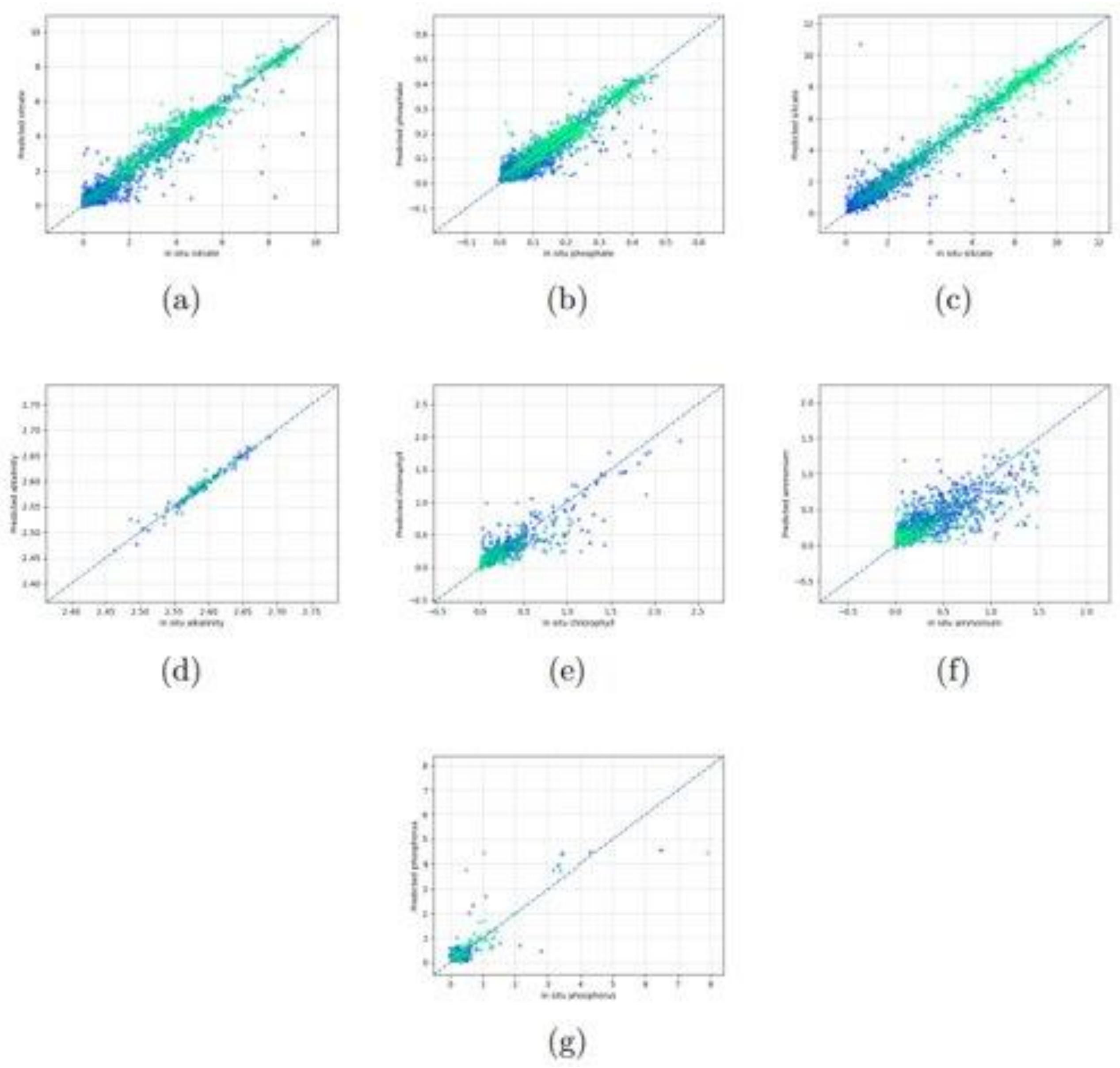
Comparison of the fitness values between current state of the art and our model. Best results are highlight with bold characters.

OGS
Istituto Nazionale
di Oceanografia
e di Geofisica
Sperimentale

Results

Comparison between the in situ measurements and the prediction's output of our model.

a) Nitrate
b) Phosphate
c) Silicate
d) Alkalinity
e) Chlorophyll-a
f) Ammonium
g) Phosphorous



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Prediction's Confidence Interval

GOAL : provide **an esteem for the uncertainty related to the estimate**, to understand *how much our prediction is reliable*.

let's introduce the concept of **confidence interval**, that is an interval statistic used to quantify the uncertainty on a prediction.

The ensembling of Neural Networks, also called **Deep Ensembles**, is a method for predictive uncertainty estimation.

OGS

Istituto Nazionale
di Oceanografia
e di Geofisica
Sperimentale

Prediction's Confidence Interval

Once given an input and after computing the ten corresponding output variables let's consider the **difference between the third quartile and the first quartile**.

The **gap** among these two quantities is an **indicator of how much our predictions is reliable**.

Practical observations performed during the analysis of the prediction confirms that small difference between quantiles corresponds to small error in the prediction of the results.

OGS
Istituto Nazionale
di Oceanografia
e di Geofisica
Sperimentale

Prediction's Confidence Interval

Comparison between the error in the measurements and the corresponding difference between the third and the first quartile.

a) Nitrate

b) Phosphate

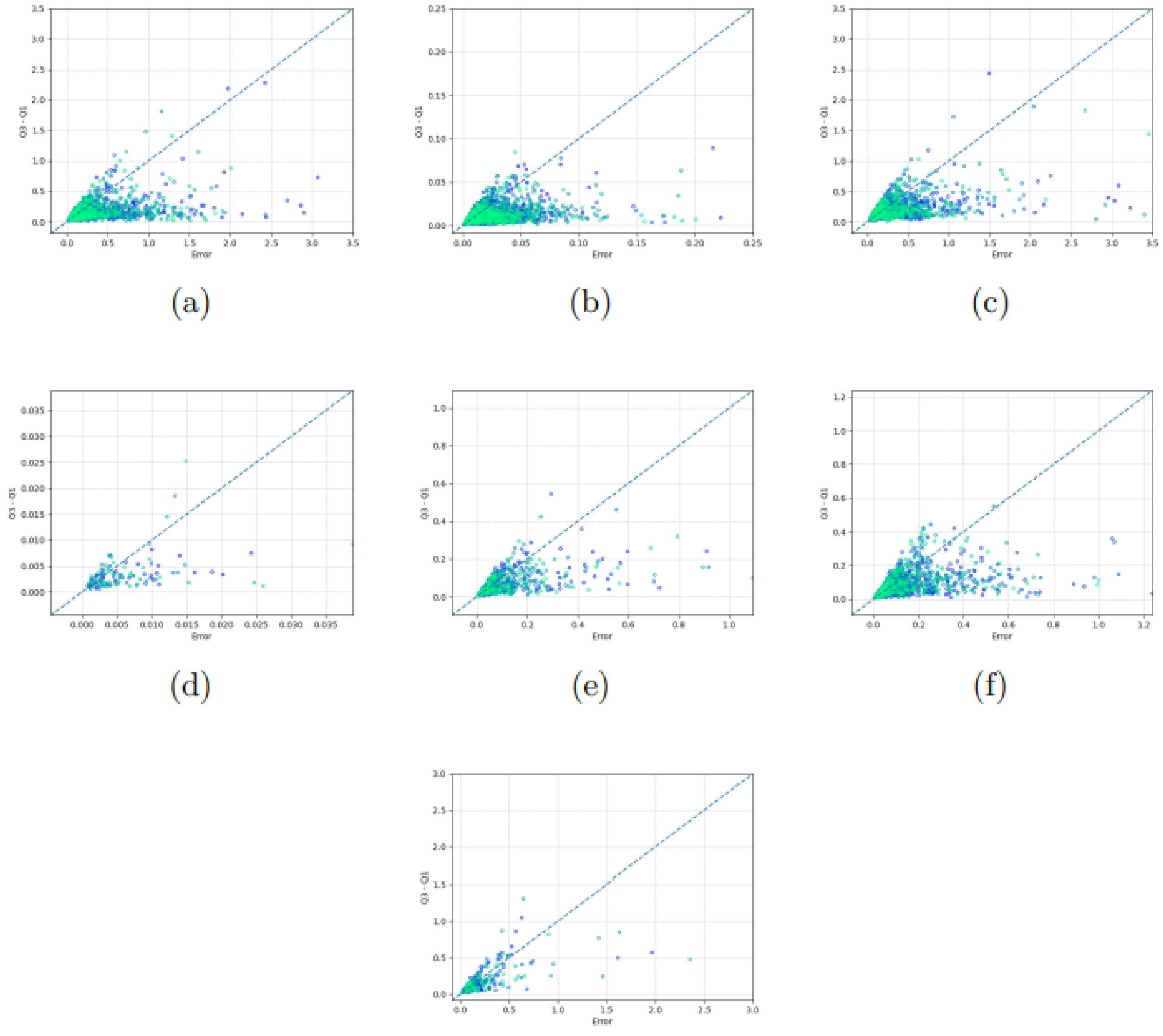
c) Silicate

d) Alkalinity

e) Chlorophyll-a

f) Ammonium

g) Phosphorous



(a) (b) (c)

(d) (e) (f)

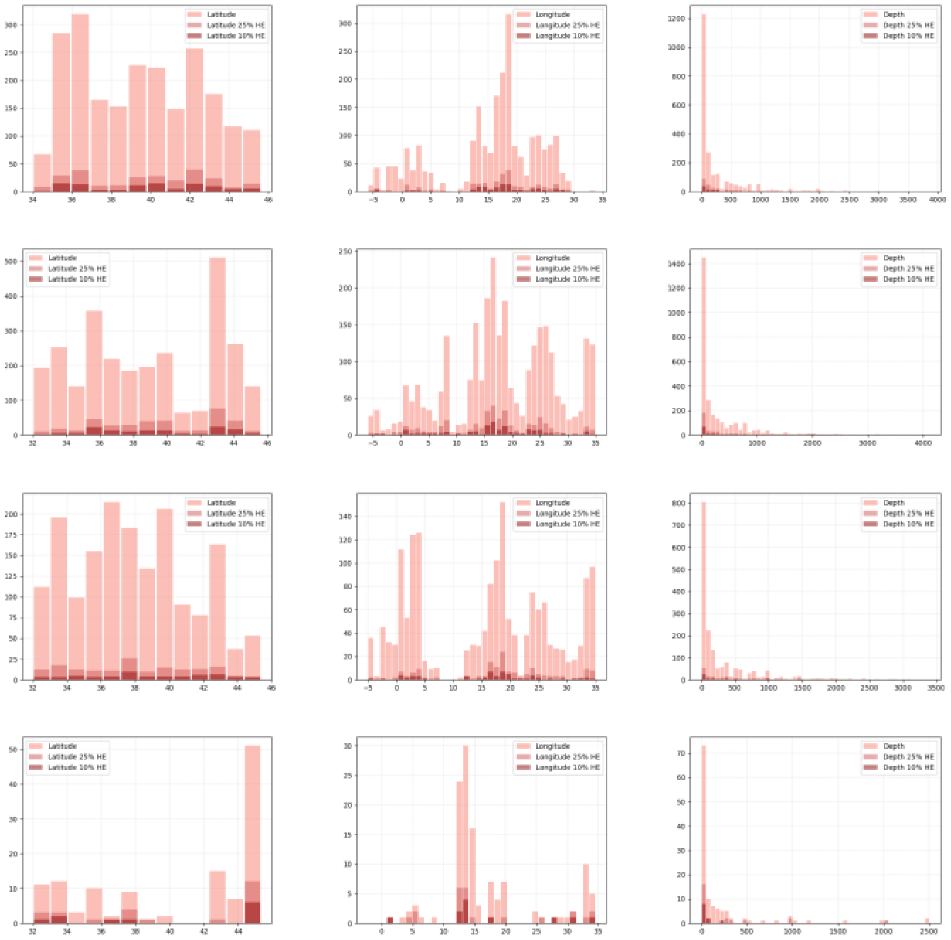
(g)

CONCLUSION

Latitude

Longitude

Depth



Nitrate

Phosphate

Silicate

Alkalinity

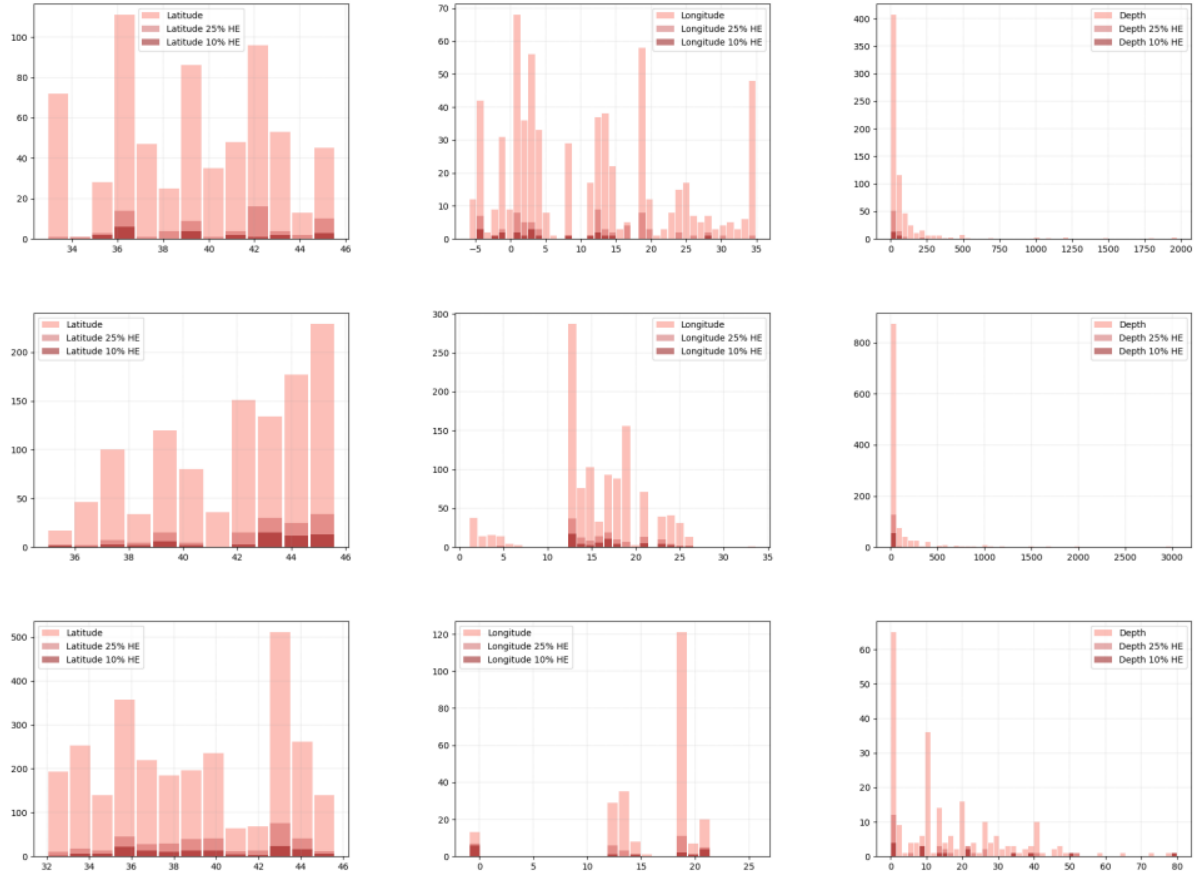
Histogram of the validation set. With red color are highlighted 10% of higher errors input data. Dark orange highlight 25% of higher errors input data

CONCLUSION

Latitude

Longitude

Depth



Chlorophyll-a

Ammonium

Phosphorous

Histogram of the validation set. With red color are highlighted 10% of higher errors input data. Dark orange highlight 25% of higher errors input data

CONCLUSION

For all the variables there is an important **decrease** for **both the fitness metrics chosen**, ensuring a significant improvement in the quality of the model.

- **The dataset** that has been considered for the training of the MLPs, that has both a major quantities of samples and a broader coverage of the Mediterranean Sea area.
- **The routine of removing boats** linked to the outliers data.

Thanks to the informations stored in the EMODnet dataset, we predict also the values for Chlorophyll-a, Ammonium and Phosphorus, unlike the deep learning techniques.