Some optimization algorithms for Ocean Water IOP retrieval
Global and Multiobjective Optimization exam 2022-2023

Riccardo Percacci

0.1 Summary

The objective is to explore some evolutionary algorithms related to the retrieval of Inherent Optical Properties (IOP) of ocean water. Two Artificial Neural Networks (aNN) were evolved for regression using NEAT's Python implementation, and a $(\lambda + \mu)$-evolution strategy was implemented for the fitting of a fourth-order polynomial function of the log band ratio.

0.2 Optical Properties of Ocean Water

Optical Properties of Water are subdivided into two classes: inherent (IOP) and apparent (AOP). IOPs are properties of the medium that do not depend on the presence of light, such as the absorption coefficient. On the other hand, AOPs, e.g. radiance, are dependent on the ambient light field. Often one wants to predict IOPs from the observation of some AOPs.
Spectral Remote-Sensing Reflectance, $R_{rs}(\lambda)$, is one of the most used products of satellite imagery. As an AOP, it is often used as a predictor. $R_{rs}(\lambda)$ is defined as the ratio between spectral water-leaving radiance and surface downwelling irradiance:

$$R_{rs}(\lambda) = \frac{L_w(\lambda)}{E_s(\lambda)}$$

Another relevant AOP is the spectral diffuse attenuation coefficient for downwelling irradiance, $K_d(\lambda)$.
The most important IOPs are the absorption and backscatter coefficients. Spectral absorption and backscatter are expressed as the sum of components:

$$a(\lambda) = a_w(\lambda) + a_{ph}(\lambda) + a_{dg}(\lambda),$$

$$b_b(\lambda) = b_{bw}(\lambda) + b_{bp}(\lambda),$$

where a_{ph} is phytoplankton absorption, a_{dg} is detritus and Gelbstoff (or CDOM - colored dissolved organic matter) absorption, and b_{bp} is particle backscatter. a_w and b_{bw}, the absorption and backscatter coefficients of pure water, are assumed to be known. Proportions of these components vary among the different water types. Phytoplankton concentration is an important marker of water type.

0.3 The data

Three data sets were gathered. The first is the Synthesized Dataset from IOCCG Report 10 [1]. For this data set, IOPs were generated using various optical parameters and models, and AOPs were modeled using the HydroLight radiative transfer numerical model with the provided IOPs. The declared purpose of this data set is to provide testing grounds for ocean color algorithms.
Two in-situ datasets were also employed, the NASA bio-Optical Marine Algorithm Data set (NOMAD) [2] and the Arctic Bio-Optical Database [3]. The former data set includes recordings from a range of ocean water types, unlike the latter, which focuses on the Arctic Ocean. From these, data was extracted according to the variables of interest, and where such data was not missing.

0.4 Algorithms

0.4.1 aNN 1

In [1], an aNN is discussed for the prediction of $a_{ph}(443)$, $a_{dg}(443)$, and $b_{bp}(443)$ values, given measurements of $R_{rs}(\lambda)$ for λ at wavelengths 411, 443, 489, 510, 560, 619, 665. We try to implement a similar algorithm via neuroevolution, using NEAT-Python.
As in [1], the quantities of interest are log-transformed. Training is performed on synthetic data, while 118 complete samples were retrieved from the in-situ datasets for validation.
By plotting in log-space the observed values of $a_{ph}(443)$, $a_{dg}(443)$, and $b_{bp}(443)$ against $R_{rs}(\lambda)$ at different wavelengths, we can observe some close to linear relations. A two-layer network should then be able to decently approximate the true model.

In order for the algorithm to find the bands that correlate most with the dependent variables, the initial connectivity of new genomes is set to partial direct, with only a 0.2 chance of any connection being present. The only activation function allowed is 'identity', while the aggregation is set to 'sum'. Bias, response and connection weights are allowed to mutate in a range from -10 to 10.
To evaluate the genomes, we activate the respective neural network, and measure its accuracy over the training data set. After 10000 generations, the following network is returned as the best:

```
Best genome:
Key: 1866019
Fitness: 614.2076676148863
Nodes:
        0 DefaultNodeGene(key=0, bias=1.606127437504259, response=-0.369814765611306, activation=identity, aggregation=sum)
        1 DefaultNodeGene(key=1, bias=-0.3787193583828017, response=-0.9548234800870625, activation=identity, aggregation=sum)
        2 DefaultNodeGene(key=2, bias=2.288628036195 2815, response=-0.7813930239493236, activation=identity, aggregation=sum)
Connections:
        DefaultConnectionGene(key=(-7, 0), weight=-2.0507932841730536, enabled=True)
        DefaultConnectionGene(key=(-7, 1), weight=-0.9310687326284474, enabled=True)
        DefaultConnectionGene(key=(-6, 2), weight=-1.3863778817824557, enabled=True)
        DefaultConnectionGene(key=(-2, 0), weight=0.42838878711490724, enabled=True)
        DefaultConnectionGene(key=(-1, 1), weight=0.804461569210138, enabled=True)
```

Four out of seven bands were used: 411, 443, 619 and 665.

0.4.2 aNN 2

Another aNN is evolved to predict $a(443)$ and $b_b(443)$, given measurements of $K_d(443)$ and $R_{rs}(\lambda)$ for λ at 411, 443, 489, 510, 560, 619, 665. This problem is inspired by an inversion algorithm presented in [4]. Again, all data is first log-transformed.
As in the previous neural network, initial connectivity is set to partial direct with 0.2 probability of any connection being present, to help select the most useful predictors. Activation functions are restricted to 'identity', aggregation to 'sum'. An analogous approach is used for evaluating the genomes.
The synthetic IOCCG data set was used for training, while validation was done through the combined in-situ data sets, with 143 full samples. This is the best network after 10000 generations:

```
Best genome:
Key: 439779
Fitness: 937.8370761650231
Nodes:
        0 DefaultNodeGene(key=0, bias=-0.31440507826474484, response=-1.8132721980306046, activation=identity, aggregation=sum)
        1 DefaultNodeGene(key=1, bias=-0.13605960893856855, response=0.30766789151531815, activation=identity, aggregation=sum)
        83938 DefaultNodeGene(key=83938, bias=2.4863965835938178, response=1.213126953024143, activation=identity, aggregation=sum)
Connections:
        DefaultConnectionGene(key=(-8, 0), weight=-0.574819880060476, enabled=True)
        DefaultConnectionGene(key=(-8, 1), weight=2.002480284682115, enabled=True)
        DefaultConnectionGene(key=(-4, 1), weight=1.9294296557995942, enabled=True)
        DefaultConnectionGene(key=(0, 83938), weight=0.05634659570723205, enabled=True)
```

$K_d(443)$ and $R_{rs}(510)$ are the only variables used as predictors.

0.4.3 $(\lambda + \mu)$−ES for chlorophyll-a concentration model tuning

Many ocean color algorithms heavily rely on the ratio (R) of R_{rs} at blue to green wavelengths [5]. The maximum band ratio (MBR) selects the ratio that gives the maximum value for R from several bands, and is here defined as

$$R = \frac{\max(R_{rs}(443), R_{rs}(490), R_{rs}(510)}{R_{rs}(555)}.$$

When observations of MBR and Chlorophyll-a concentrations are plotted in log-space, the shape of the data is sigmoid. Empirical algorithms express this functional relation as a polynomial with respect to the log-transformed quantities. Thus, we aim at fitting a model

$$\log_{10}[\text{Chl}] = a + b \log_{10} R + c(\log_{10} R)^2 + d(\log_{10} R)^3,$$

This model was fitted first using synthetic data, then the combined in-situ data sets as training data. For the synthetic data set, [Chl] was obtained by using the formula

$$a_{ph}(440) = 0.5[\text{Chl}]^{0.626}.$$

As the formula above was used to generate a_{ph} values for the dataset, this does not introduce any additional bias to the model. 1349 complete samples were retrieved from in-situ data.
The target function to minimize during evolution is

$$f(a,b,c,d) = \sum_{i=1}^{N} (y_i - a - bx_i - cx_i^2 - dx_i^3)^2,$$

where N is the number of sample points, and x_i ad y_i are the maximum log band ratio and chlorophyll-a log-concentration (in [mg m^{-3}]) for sample i.
The evolution was run for 1000 generations in both cases, and the best parameteres were retrieved for testing.

0.5 Results

0.5.1 aNN 1

We test the best network on training data, and obtain 0.94, 0.98, 0.98 as the R^2 (coefficient of determination) of the best linear fit, with relation to the predicted variables. Fig. 1 gives a graphical view of regression results on training data.

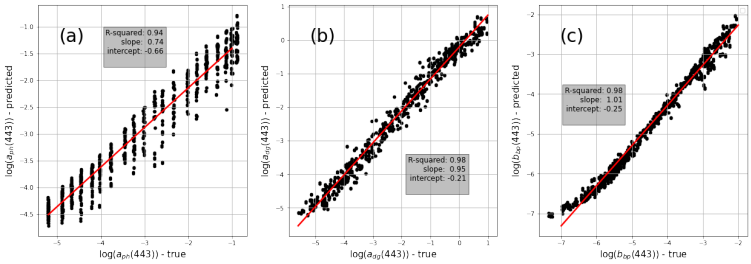


Figure 1: Training error for: (a) - $\log a_{ph}(443)$, (b) - $\log a_{dg}(443)$, (c) - $\log b_{bp}(443)$.

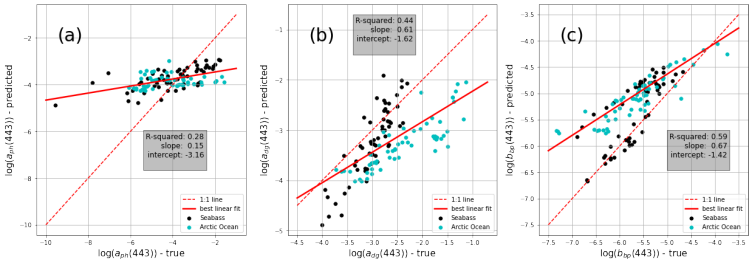
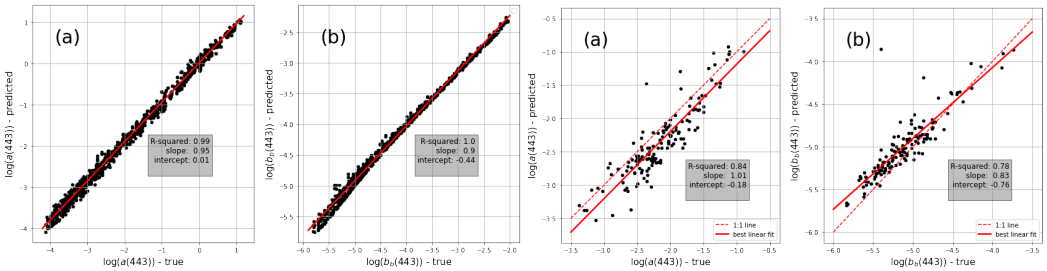


Figure 2: Validation error for: (a) - $\log a_{ph}(443)$, (b) - $\log a_{dg}(443)$, (c) - $\log b_{bp}(443)$. Black dots are Seabass data points, while in cyan are Arctic Ocean data points. Seabass data seems to agree more with the model.

The model didn't perform too well on in-situ data, especially in the case of samples from the Arctic Ocean data set (Fig. 2a).
Arctic Ocean waters differ sigificantly from most ocean waters, also due to the lower chlorophyll-a concentration, which alters absorption and backscatter coefficients [6]. The Seabass data set is more generic and covers a wider range of water types, which means the model trained on synthetic data performs better on it.
The results of in-situ validation testify to the difficulties and errors encountered in modeling a wide range of water types. To overcome this, often a model is developed with a specific water type in mind.
A few outliers appear among the $\log a_{ph}(443)$ Seabass samples, in the lower range. The true values of these samples fall outside the range of values encountered in training, leading to large errors.

0.5.2 aNN 2

By looking at Figure (Figure 3a), we see the model fitting the training data quite well. Validation on in-situ data (Figure 3b) also shows good results, with R^2 coefficients of 0.84 and 0.78 for best linear fits. As could be expected,



(a) Training Error (b) Validation error

Figure 3: aNN performance on training data and validation data

total absorption and total backscatter coefficients are easier to model than absorption due to phytoplankton or gelbstoff, or particle backscatter, using simple neural networks.

0.5.3 $(\lambda + \mu)$−ES

The first model was trained on synthetic data. It fit the data well, and performs quite well on in-situ data (R^2 = 0.81, slope= 0.92) (Fig. 4).

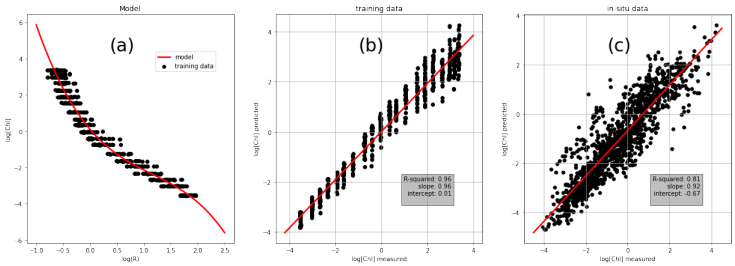


Figure 4: Training on synthetic data

We also train the same model on in-situ data, and plot it with the previous model, evaluated on in-situ data (Fig. 5).

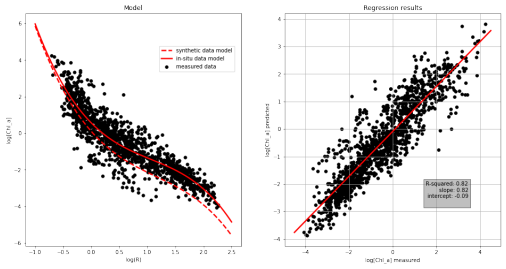


Figure 5: Training on in-situ data

We can see that the model trained on synthetic data slightly underestimates in-situ values of $\log_{10}$[Chl].

0.6 Conclusions

Two artificial neural networks were evolved using NEAT-Python, to predict inherent optical properties of water. Their training was done on a synthetic data set, while in-situ data was collected for validation. Results showed that prediction of total absorption and backscattering is easier than their components (e.g. absorption due to phytoplankton or particle backscatter). This is because the composition of ocean waters has a range of variability, most importantly of phytoplankton concentration, which makes modelling of certain IOPs more restricted

geographically. In any case, it seems that artificial neural networks represent a valid option for algorithm development in this domain.
The problem of tuning four parameters for the fitting of a fourth order polynomial to given data represented an easy target for a $(\lambda + \mu)$-ES approach. In contrast to the two neuroevolution examples above, results were obtained much faster, and performed well when validated on in-situ data.

Bibliography

[1] IOCCG (2006). Remote Sensing of Inherent Optical Properties: Fundamentals, Tests of Algorithms, and Applications. Lee, Z.-P. (ed.), Reports of the International Ocean-Colour Coordinating Group, No. 5, IOCCG, Dartmouth, Canada

[2] Werdell, P.J. and S.W. Bailey , 2005: An improved bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote Sensing of Environment* , 98(1), 122-140.

[3] Lewis, Kate; van Dijken, Gert; Arrigo, Kevin (2020). Bio-optical Database of the Arctic Ocean [Dataset]. Dryad. https://doi.org/10.5061/dryad.cnp5hqc17

[4] Loisel, H., Stramski, D., Dessailly, D., Jamet, C., Li, L., Reynolds, R. A. (2018). An inverse model for estimating the optical absorption and backscattering coefficients of seawater from remote-sensing reflectance over a broad range of oceanic and coastal marine environments. *Journal of Geophysical Research: Oceans*, 123, 2141–2171. https://doi.org/10.1002/2017JC013632

[5] O'Reilly, J. E., Werdell, P. J. (2019). Chlorophyll algorithms for ocean color sensors - OC4, OC5 & OC6. *Remote Sensing of Environment*, 229, 32-47. ISSN 0034-4257. https://doi.org/10.1016/j.rse.2019.04.021.

[6] K.M. Lewis, B.G. Mitchell, G.L. van Dijken, K.R. Arrigo (2016). Regional chlorophyll a algorithms in the Arctic Ocean and their effect on satellite-derived primary production estimates. *Deep Sea Research Part II: Topical Studies in Oceanography*, 130, 14-27. ISSN 0967-0645. https://doi.org/10.1016/j.dsr2.2016.04.020.