ML Engineer Nanodegree Capstone Project

Machine learning estimates of anthropogenic CO_2

Background

Since the beginning of the industrial revolution (1860), anthropogenic activities have increased the atmospheric carbon dioxide (CO_2) from 280 ppm to over 400 ppm [6]. The global ocean has sequestered roughly a third of this anthropogenic CO_2 (C_{ant}, Fig.1), limiting the impacts on the global climate [3] such as the increase in the Earth's mean temperature. However, C_{ant} can only be estimated indirectly in the ocean with an uncertainty of ±20%. This uncertainty reduces our understanding of the processes associated with the C_{ant} cycle and the reliability of future predictions.

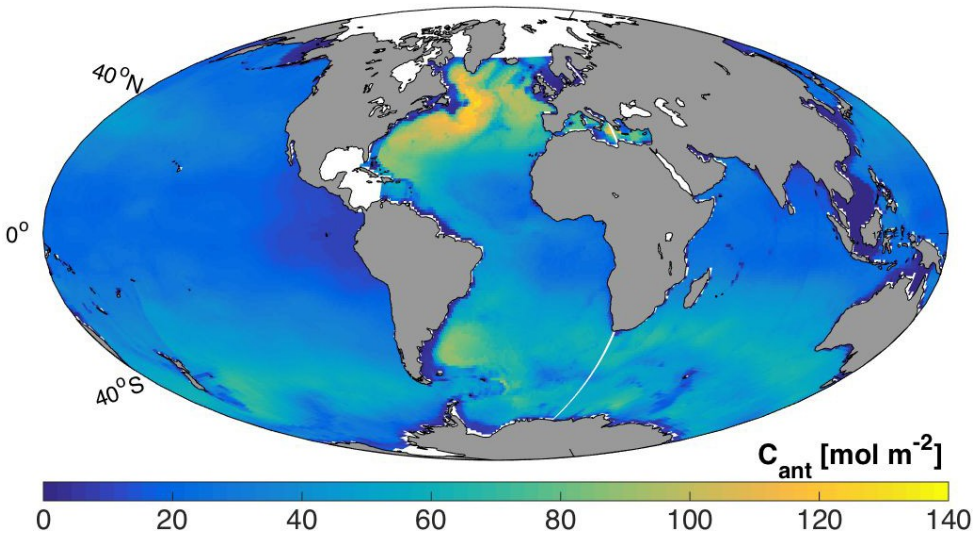


Fig. 1: *Global ocean anthropogenic carbon (C_{ant}) column inventory. Data taken from the GLODAPv2 climatology [4], where they were estimated using the Transit-Time Distribution (TTD) method based on dichlorodifluoromethane (CFC-12) observations [7].*

Problem

C_{ant} estimates cannot be measured directly in the ocean being instead approximated using correlations with measurable variables (e.g. inorganic nutrients).

(1) The most commonly adopted methods for C_{ant} estimates [2, 7] use Redfield ratios to isolate C_{ant} from the ocean total dissolved inorganic carbon (tCO2) measurements. Redfield ratios are ratios of inorganic nutrients used to infer biological activities in the ocean [5]. Redfield ratios vary over space and time, based on the in-situ species.

(2) The transit-time distribution (TTD) technique [7] assumes that chlorofluorocarbons (CFCs) can be used as proxies for C_{ant} estimates. However, CFCs have been released in the atmosphere approximately a century after the release of C_{ant}: CFCs and C_{ant} distributions are comparable only in the younger water masses of the ocean.

(3) The extended multi-linear regression (eMLR) technique [1] estimates C_{ant} changes over time assuming that correlations between C_{ant} and predictors are constant. This assumption is unlikely, being the ocean in continuous variation.

The effects of those three assumptions lead to an uncertainty in the C_{ant} estimates of approximately ±20% across different methods. This result reduces the understanding of the C_{ant} cycle and the possibility of predicting its future.

Motivation

As chemical oceanographer, I have spent my Ph.D analyzing the most common methodologies (e.g. TTD, eMLR) used to estimate C_{ant} . Although the combination of multiple approaches reduces the uncertainty on the C_{ant} estimates, this value cannot be reduced below ±10% with the current approaches: the most commonly used methods for C_{ant} estimates rely on a priori assumptions of the ocean system (e.g. Redfield ratio values). Conversely, the most common machine learning algorithms learn from data, overcoming the necessity of a priori assumptions. So, I am keen to test the potentialities of machine learning algorithms to improve current C_{ant} estimates and therefore better understanding the processes that control their variations.

Approach

Introduction

The majority of the algorithms currently used in machine learning does not require prior assumptions but learns from data. This overcomes the strongest limit in the methods currently used to estimate oceanic C_{ant}.
My analysis will be performing into three steps. Initially, the GLODAPv2 [4] data will be explored by using box and scatter plots together with correlation matrices (see next section). This provides a data overview and identifies features correlations and specific patterns. As a second step, all of the most commonly used regressors (linear regression, decision tree, random forest, neural network, and lightGBM, see next section) will be used to predict C_{ant} and their performances will be compared with the TTD and benchmark model described below. Finally, the error of the most performing models will be explored and discussed together with future steps and potential improvements.

Algorithms

The *linear regression* is a statistical approach used to model the relation between a scalar variable of interest and one or more predictors. A linear regression [10] can be summarized as $y = a + bx$, where y is the variable of interest, a is the constant term, and b is the generic coefficient(s) of the predictor(s) x. In practical terms, a common approach to linear regression is the Scikit learn module [14] used in this analysis.

The *ridge* [13] and *lasso* [12] *regressions* are statistical methods developed to overcome linear regression challenges. The ridge approach [11] is useful when data include an excessive amount of variables with respect to the data collected for each or include highly correlated predictors, such as in the GLODAPv2 dataset. The least absolute shrinkage and selector operator (lasso) regression performs the selection and regularization of the predictors used in the linear regression. The lasso approach guarantees comparable effects of each predictor.

The *decision tree* [15] is a machine-learning algorithm based on a decision tree clustering system that predicts values or classes of a variable of interest. The decision tree is normally used to visually represent a decision making approach but it can also be used to predict the amounts of a scalar variable, such as C_{ant}.

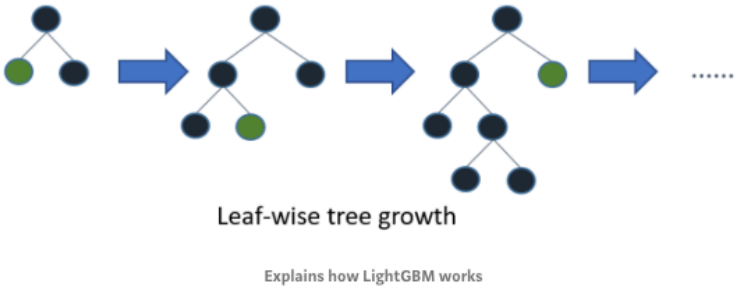
The *random forest* algorithm [16] is an evolution of the decision tree that uses multiple combinations of multiple decision trees developed horizontally. The use of multiple decision trees captures a larger fraction of the variable variance and reduces the error on the final estimate, but enlarges the computer power and time required.

The *neural network* algorithm [17] is more complex than the previously described methods. Artificial neural networks are systems designed to be similar to the biological network of neurons and able to learn from data without requiring specific coded instructions. More in detail, each neuron of the network receive a signal, process it accordingly to specifically designed functions that determine the degree of activation of the neuron, and propagates the signal to the following neuron or to the end of the system. I use the multi-layer perceptron (MLP) implementation in Scikit learn [18]. In formal terms, a function of activation for a neural network can be:

$$f(y_i) = \begin{cases} y_i, & \text{if } y_i > 0 \\ a_i y_i, & \text{if } y_i \leq 0 \end{cases}.$$

This activation function shows that each neuron of the network would act differently depending on the information carried by the incoming signal. Only if this signal reaches a pre-determined threshold, the neuron activates and propagates it further.

The *lightGBM* algorithm is ‘a gradient boosting framework that uses tree based learning algorithm’ [19]. This relatively new approach is similar conceptually to the decision tree and random forest algorithms but it grows the trees vertically instead of horizontally (see diagram below). In another words, the lightGBM approach includes multiple trees each of which has a different combination of leaves combinations and not only new static trees are added to the analysis.



Explains how LightGBM works

How other boosting algorithm works

The two main reasons why the lighGBM is getting so popular in the machine-learning word and so it is used here is that the vertical increase in the tree approach allows to work more efficiently and faster, and the structure of the algorithm has been built to focus on the result accuracy.

Dataset and Input

Introduction

As dataset, I will use the freely accessible data from the second version of the global ocean data analysis project (GLODAPv2, https://www.glodap.info/). This data include measures of the most common oceanography variables, such as dissolved inorganic carbon, total alkalinity, inorganic nutrients, and a C_{ant} estimate [4]. C_{ant} is estimating using the transit-time distribution (TTD [7]) technique and it will be considered here as the baseline for the proposed machine learning models. C_{ant} can also be quantified in the GLODAPv2 data as difference between industrial and pre-industrial dissolved inorganic carbon (tCO2), both of which are given in the GLODAPv2 dataset. I use this second estimate as C_{ant} reference for all of the models results.
The GLODAPv2 dataset provides climatology for the year 2006, which allows applying machine learning algorithms globally and testing their performances with respect to a homogeneous set of data. Also, the GLODAPv2 data include a detailed quality check, which reduces the necessity of data exploration and analysis.

Dataset

The dataset description is given on a random subset including 0.05% (520 records) of the 2006 GLODAPv2 climatology dataset (1,039,941 records). This data reduction improves the analysis velocity and the laptop performances, being significant thanks to the randomness of the data selection.
The GLODAPv2 climatology data are shown in Fig.2 as scatter plots and kernel density estimations [9] (kde, continuous lines). The formers identify correlations between features (TTD anthropogenic carbon (Cant_ttd), inorganic nitrates (Nitr), dissolved oxygen (Oxyg), pH measured at 25°C (pH25), inorganic phosphate (Phos), salinity (Sali), inorganic silicate (Sili), total alkalinity (Talk), tCO2, temperature (Temp), and reference C_{ant} (Cant_ref)), while the latters show the probability distribution of the data density.
As expected, all of fields shown in Fig.2 are correlated with C_{ant}. The only exceptions are Sali and Talk, but this effect is due to a small amount of high and low values in both distributions, whereas a study focused on the second and third quartiles would have shown higher correlations for both features.
The kde distributions show a similar result with the majority of the data lying around the averages for Sali and Talk. Other fields are more homogeneously distributed with peaks towards the kde extremes (e.g. Phos).

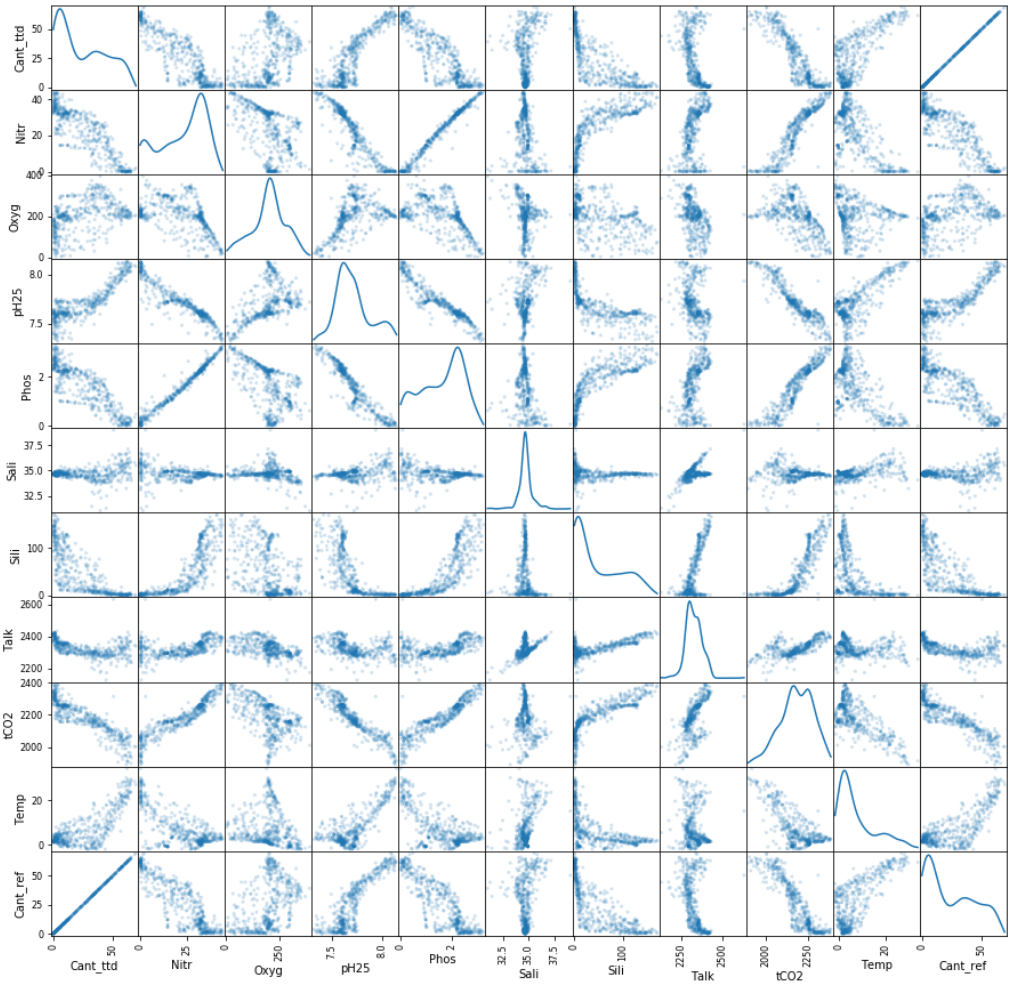


Fig. 2: *Scatter matrix of ocean data (TTD anthropogenic carbon (Cant_ttd), inorganic nitrate (Nitr), dissolved oxygen (Oxyg), pH measured at 25°C (pH25), inorganic phosphate (Phos), salinity (Sali), inorganic silicate (Sili), total alkalinity (Talk), dissolved inorganic carbon (tCO2), temperature (Temp), and reference C_{ant} (Cant_ref)). X and y axes are identical with the scatter plots showing the correlations between features and the continuous lines showing the kernel density estimation [9] of each feature. Data units are not shown for simplicity.*

Fig.3 shows the GLODAPv2 climatology data distributions after a calibration that forces them to vary between -1 and 1. This approach allows comparing the feature distributions and identifying common patterns. For instance, all data are showing a greater density around the averages, which is zero in the transformed coordinates of the graph. Some distributions are normally distributed (e.g. Nitr), while others are more skewed towards the extremes (e.g. Sili).

Finally, Fig.4 shows the correlations between pairs of oceanographic fields. All of the features analyzed are correlated with C_{ant}. This is true also for Sali and Talk, with respective values of 0.20 and 0.40.

Omega A and C are also provided in the GLODAPv2 climatology but are not used in this analysis as they are not common in the oceanographic data. The in situ measurement of pH is also given in the dataset but it is ignored in the present analysis as it is highly correlated with the pH25 distribution.

Fig. 3: *Box and scatter plots showing calibrated distributions of oceanographic data. Calibration applied by removing the distribution average and dividing by the distribution range (max - min).*

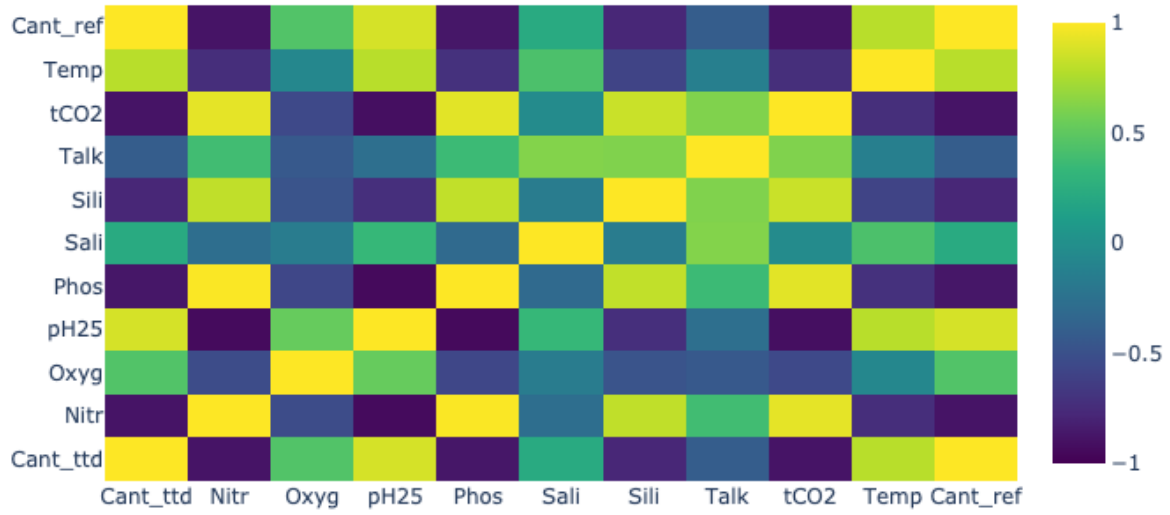


Fig. 4: *Correlation matrix of oceanographic data based on the GLODAPv2 climatology data.*

Benchmark model

The transit-time distribution technique [7] assumes that the anthropogenic CO_2 can be treated as a transient tracer and so its concentration can be considered uninfluenced by biological cycles when entered the ocean. If so, other transient tracers, such as CFCs, can be used as proxies for it. This relies on complicated equations and would require a longer explanation, but for the analysis in this project, I summarize it as if changes in CFCs can approximate C_{ant} variations under few assumptions not always fulfilled. The benchmark model should be able to challenge the TTD approach.
As a benchmark model, I will consider a relatively simple linear regression based on all available fields. This model results will be compared with the Cant_ref, used in this study as the reference. More complicated models would also be trained and compared with the TTD, the benchmark, and the Cant_ref C_{ant} values.

Evaluation metric

The choice of the evaluation key performance indicator (KPI) is probably the most challenging aspect of the project. As mentioned, C_{ant} cannot be measured directly in the ocean but it is estimated indirectly from other variables. I will assume the ref C_{ant} estimates to be the “truth” towards which all other methods will be compared. The comparison will be based on the calculation of the R squared (R^2), the root mean standard error (RMSE), and the main absolute error (MAE).
The R^2, RMSE, and MAE are chosen because they are the most used calculations. Among them, however, the R^2 will receive the least consideration, as its units are not the same as the predictions. Between RMSE and MAE, the latter will receive greater attention because it penalizes less the discrepancies among predictions and measurements.

Analysis

Based on the analyses conducted in the dataset introduction section, I choose the benchmark model to be a simple linear regression based on nine features: Nitr, Oxyg, pH25, Phos, Sali, Sili, Talk, tCO2, and Temp. No model parameters have been selected a priori to avoid the risk of link the model to the dataset used for the train and hence generate overfitting that may increase the uncertainty on the final C_{ant} estimates.
The initial dataset has been randomly reduced to 10% (103,994 records) to improve the speed of the analysis and the laptop performance. Of this initial data, 20% (20815 records) have been stored in the test and 80% (83257 records) in the train datasets. As per common practice, the model has been trained on the train dataset and the results compared with the Cant_ref on the test dataset (see Fig.5).

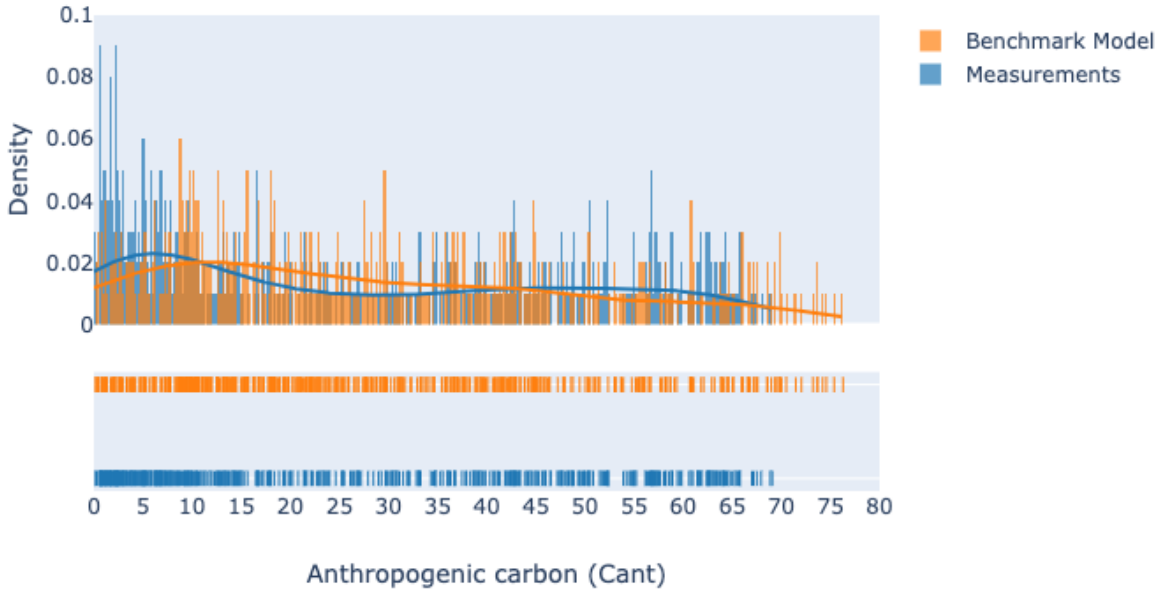


Fig. 5: Density distributions of predicted and measured anthropogenic carbon (C_{ant}). Data are taken from the GLODAPv2 climatology for the year 2006 and predictions are estimated with a linear regression, which is used hereafter as benchmark model. Only the initial 500 rows of the test dataset are plotted for visibility.

Fig.5 shows the benchmark model C_{ant} predictions in comparison with the Cant_ref, which are named 'measurements' in the figure. In the upper panel histograms have been plotted for both distributions together with the associated kde. In the lower panel, the histograms are shown from the above, highlighting the areas of greater density of data. Only the initial 500 rows of the benchmark model results and measurements are reported to improve the graph visibility.
Overall, Fig.5 shows consistency between the benchmark model results and the C_{ant} measurements. Differences exist in the low estimates, with the density peak being at 15 for the benchmark model while it is at 6 for the reference. Also, differences exist on the other extreme of the distributions, with the benchmark model quantifying C_{ant} values to a maximum of 77, while the measurements maximum is 69.
The benchmark model is a good baseline for the C_{ant} distributions that hopefully will be challenged by other models hereafter (see Tab.1).

Model Name	R^2	RMSE	MAE
Linear Regression	0.91	6.30	4.90
Ridge Regression	0.91	6.30	4.90
Lasso Regression	0.90	5.53	5.12
Decision Tree	0.98	2.66	1.55
Random Forest	0.99	**1.97**	**1.23**
Neural Network	0.91	6.42	4.77
LightGBM	0.99	2.03	1.38

Tab. 1: Performances of machine learning algorithms (Model Name) based on squared R (R^2), root mean square error (RMSE), and main absolute error (MAE). Those key performance indicators are calculated by comparing the algorithm C_{ant} predictions with the GLODAPv2 reference estimate.

Having defined and explored the model baseline versus the benchmark in the previous paragraph, I am now ready to train six additional algorithms and explore their performances in Tab.1 and Fig.6.

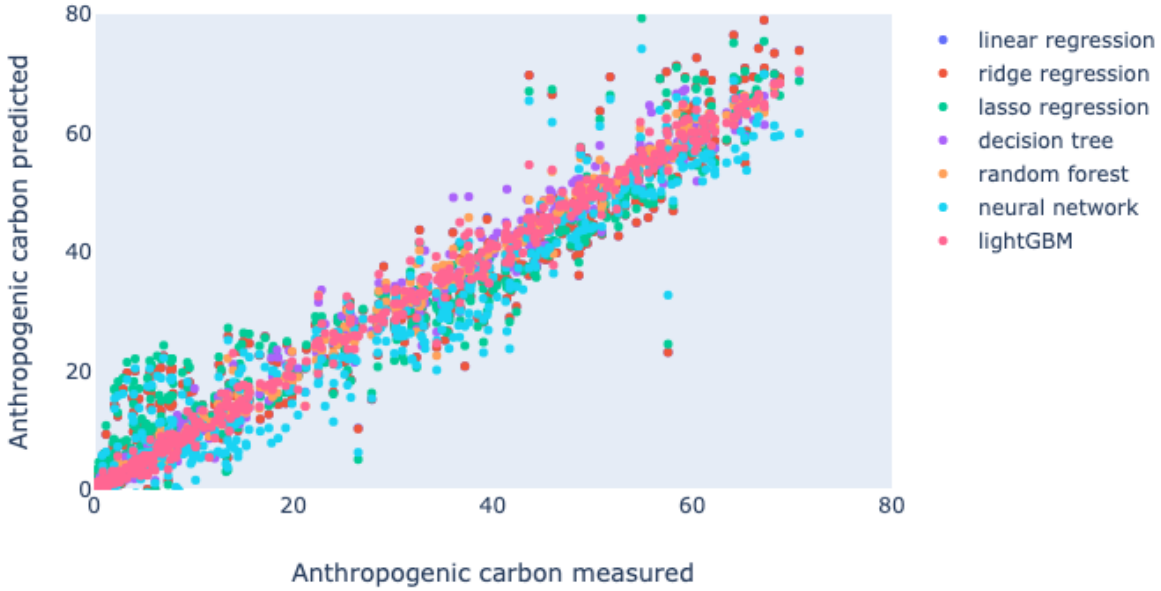


Fig. 6: *Scatter plots of anthropogenic carbon predictions versus anthropogenic carbon measurements. Data are quantified for the GLODAPv2 climatology by using seven machine-learning algorithms, which are specified in the graph legend. The linear regression is considered as benchmark model.*

Initially, I explored the use of alternatives to the linear regression, such as the Lasso and Ridge approaches. Both are still relatively easy algorithms and would have been a great improvement for the benchmark model. However, as show in Tab.1, the KPIs are identical between the benchmark model and the Ridge regression, hence showing no improvements. The Lasso approach seems to better perform in terms of RMSE, but the R^2 is lower than the benchmark model one and the MAE is greater, identifying a decrease in the predictions reliability.
Having analyzed the linear alternatives and found no particular improvements, I decided to explore more complex algorithms, such as the decision tree, the random forest, the neural network, and the lightGBM. For all, I kept the same test and train data used for the linear models in order to guarantee comparability.
As shown in Tab.1, the random forest returns the lowest RMSE and MAE, being then the most promising among the studied algorithms. Those values are comparable with the LightGBM results, which however better performs in term of memory usage and speed. As a result, I will consider the Random Forest and the LightGBM as the best algorithms for our analysis.
A similar conclusion can be drawn from Fig.6 where the Random Forest and LightGBM predictions have the strongest correlations with measurements.

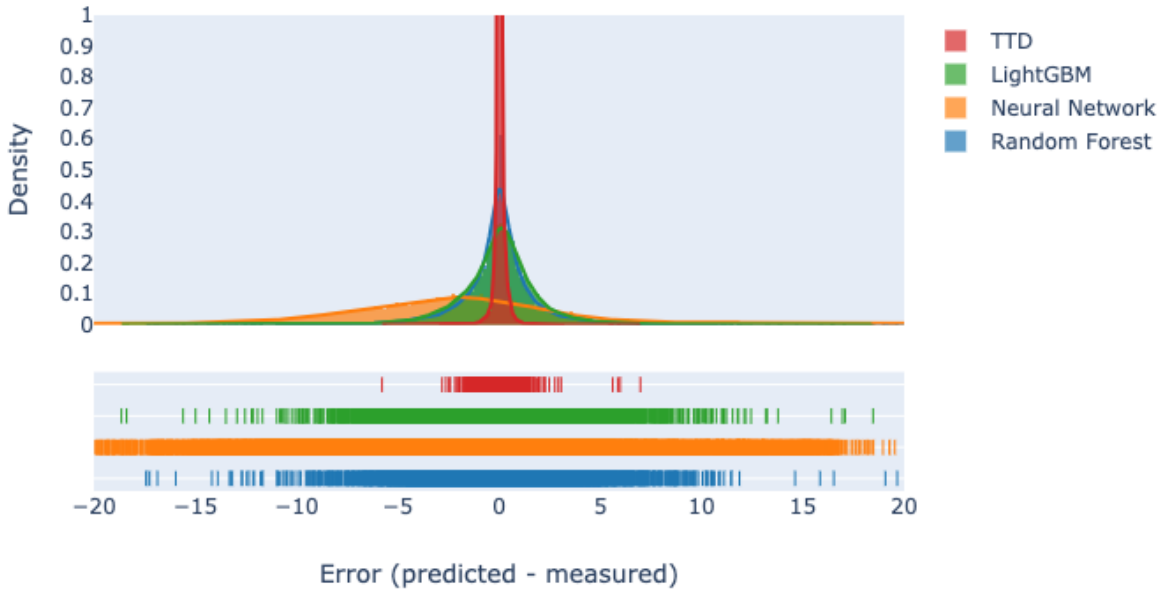


Fig. 7: *Density distributions of differences (errors) between predicted and measured anthropogenic carbon. Data are taken from the GLODAPv2 climatology for the year 2006 and estimated with the LightGBM, neural network MLP, random forest, and TTD models.*

Having identified the random forest and lightGBM algorithms as the most performing, their errors are calculated as differences with the reference and compared with the TTD and neural network (third most performing) in Fig.7. The neural network errors are spread across the entire range with a peak at -3, showing a consistent underestimation of the real C_{ant}. The neural network performance would have improved if additional data would have provided to the model in the training stage, but I have been limited by the laptop performances. The random forest and the lightGBM results are comparable and included between -10 and 10, which is approximately an uncertainty of ±13%. This is clearly greater than the TTD error, which is overall included between -3 and +3 or ±4%. However, the TTD method relies on a series of complex assumptions that could decrease its applicability, as discussed in the conclusion section.

Conclusions and next steps

An initial exploration of the machine learning possibilities to predict C_{ant} is given here. As notice in the analysis section, the TTD might seem performing better than any machine learning algorithms, but it includes some considerations that might revert the initial result.

(1) The TTD approach is based on a series of challenging assumptions, such as the absence of biological influence on the oceanic C_{ant}, while the machine learning algorithms have been left with default values on purpose to avoid overfitting.

(2) The studied machine learning algorithms showed evidences of improvements when additional data are provided, especially the neural network, random forest, and lightGBM. In this analysis, I have trained the models on a reduced fraction (10%) of the available dataset. More data can be easily provided with a more powerful machine.

(3) The analyses conducted here are based on climatological data, including no variations over time. This challenge can be overcome by using the GLODAPv2 cruise data and changes over time can be explored. It is highly likely to presume that CFCs will be available only for a small amount of cruises and so the TTD would perform on a limited fraction of the available dataset. The machine learning algorithms developed in this analysis would perform on a larger dataset providing a better description of the oceanic C_{ant}.

As a future step, I would suggest the use of the atmospheric CO_2 measurements freely provided by the US national oceanic and atmospheric administration (NOAA) to test the performances over time of the machine learning models based on the cruise data.

References

[1] Friis, K., Körtzinger, A., Pätsch, J., and Wallace, D.W.R.: On the temporal increase of anthropogenic CO_2 in the subpolar North Atlantic, Deep Sea Res. I, 52, 681-698, 2005.

[2] Gruber, N., Sarmiento, J.L., and Stocker, T.F.: An improved method for detecting anthropogenic CO_2 in the oceans, Glob. Biogeochem. Cycles, 10, 809-837, 1996.

[3] Khatiwala, S., Primeau, F., and Hall, T.: Reconstruction of the history of anthropogenic CO_2 concentrations in the ocean, Nature, 462, 346-349, 2009.

[4] S.K. Lauvset et al. A new global interior ocean mapped climatology: the 1◦ x 1◦ GLODAP version 2. Earth Syst. Sci. Data, 8, 2016. doi: 10.5194/essd-8-325-2016.

[5] Redfield, A.C.: On the proportions of organic derivations in seawater and their relation to the composition of Plankton. In J. Johnstone memorial, Liverpool University press, 176-192, 1934.

[6] Sabine, C.L., Feely, R.A., Gruber, N., Key, R.M., Lee, K., Bullister, J.L., Wanninkhof, R., Wong, C.S., Wallace, D.W.R., Tillbrock, B., Millero, F.J., Peng, T.-H., Kozyr, A., Ono, T., and Rios, A.F.: The Oceanic Sink for Anthropogenic CO_2, Science, 305, 367-371, 2004.

[7] Waugh, D.W., Haine, T.W.N., and Hall, T.M.: Transport times and anthropogenic carbon in the subpolar North Atlantic Ocean, Deep Sea Res. I, 51, 1475-1491, 2004.

[8] https://www.kdnuggets.com/2018/04/right-metric-evaluating-machine-learning-models-1.html

[9] http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/

[10] https://en.wikipedia.org/wiki/Linear_regression

[11] https://www.statisticshowto.datasciencecentral.com/ridge-regression/

[12] https://en.wikipedia.org/wiki/Lasso_(statistics)

[13] https://en.wikipedia.org/wiki/Regularized_least_squares#Ridge_regression

[14] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[15] https://en.wikipedia.org/wiki/Decision_tree_learning

[16] https://en.wikipedia.org/wiki/Random_forest

[17] https://en.wikipedia.org/wiki/Artificial_neural_network

[18] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

[19] https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc