Forecasting of Arctic Phytoplankton Abundance using Remotely Sensed Data and Machine Learning

CS229, Fall 2017

Cheenar Banerjee, Sierra Kaplan-Nelson
{cheenarb, sierrakn}@cs.stanford.edu

***Abstract*—We use various regression techniques to forecast chlorophyll a levels in the Beaufort Sea in collaboration with the Center for Ocean Solutions at Stanford. We use a dataset of remotely sensed data from 2002-2015. We experiment with weighted and unweighted regression models, and we explore the predictive value of several different features. Through our model and feature-related experiments, we explore the spatiotemporal aspects of our data. We experiment with regression models that are weighted by time, by location, and by a combination of time and location. We find that for linear and L-2 regularized linear regression, the time and space weighting improves its overall mean square error and R^2 values. However, weighting does not improve our best baseline model, which is a random forest model. We also find that features vary widely from year to year, but overall, for this particular dataset, location-based features seem to be the most predictive.**

I. INTRODUCTION

Marine satellite and in-situ monitoring can help managers locate key biological productivity hot spots in the Arctic. These hot spots form the basis of areas rich in fisheries and marine resources. Knowing the locations of key ecosystem components in the Arctic will help inform the regulation of expansion of human activities, such as oil and gas development and shipping. As part of the Center for Ocean Solutions' Catalyst project, we applied machine learning techniques to data collected from the Arctic Beaufort Sea to investigate the potential of using remotely sensed and other data to predict chlorophyll a concentrations, as well as understanding what features of the natural environment are most predictive of biological productivity. Chlorophyll a concentration is a heuristic for phytoplankton activity. Phytoplankton is at the bottom of the food chain, thus biological productivity can be approximated through phytoplankton activity. This is the first time machine learning techniques have been applied to these data, and our results will help inform future research in the area. In this paper, we mainly explore weighted and unweighted regression models to predict future chlorophyll a concentrations.

II. RELATED WORK

Machine learning algorithms including linear regression have been applied to other remote sensing data in the Baltic Sea to predict macrophyte and invertebrate species cover. Kotta et al. found that a combination of machine learning and statistical modeling is effective at predicting biological concentrations without over-fitting and without the need for prior elimination of outliers. Their results showed that boosted regression trees were the best model for the data. [1]. Using a different set of features, Ye and Cai constructed a recurrent neural network to forecast cholorphyll a levels 0 days and 7 days ahead in the Xiangxi Bay. They were able to achieve moderate success, indicating that chlorophyll a prediction is a tractable problem. [2]

Elattar et al. used a modified version of locally weighted linear regression to forecast electric load. They modified support vector regression to incorporate local weights, and showed that it acts as a successful model for electric load forecasting. [3] Sun et al. also experienced success when using a locally weighted linear regression model for short-term traffic forecasting. [4] Haack et al. and Camps-Valls et al. showed that regression techniques can be successfully applied to remotely sensed data. [5] [6]

III. DATASETS AND FEATURES

Our dataset contains remotely sensed data collected from the Arctic Beaufort Sea in the years 2002-2015. All data is presented in eight-day time step aggregates. The satellite images are represented as pixels; each data point represents one pixel and contains the following features: day of year start, day of year end, latitude coordinate, longitude coordinate, water depth (m), distance to land (m), sea ice concentration (proportion of image covered by ice), sea surface temperature (degrees Celsius), mean cloud cover (percentage), modeled mean and maximum photo synthetically active radiation (PAR) near the sea surface (uEin/m^2/s), hours of daylight, sea level anomaly (m, a measure related to ocean currents), and the corresponding chlorophyll a concentration (mg/m^3). Figure 1 shows a heat map of chlorophyll a concentrations across the area of interest for a given eight-day time step in the data. The data varies along a spatiotemporal axis – Figure 1 shows how chlorophyll a concentration varies by location, especially by location relative to the coastline. Furthermore, phytoplankton activity varies with seasonal patterns. We restrict our experiments to focus on data between early July and late September, the time of year where our data was most consistently collected and when phytoplankton activity is at its peak.

We train our models on the data from 2002-2014 (1080268 data points), and test on the data from 2015 (46298 data points). We choose to test on the latter portion of the data, rather than randomly selecting a training and testing set, due to the temporal variation we expect to see in future years.

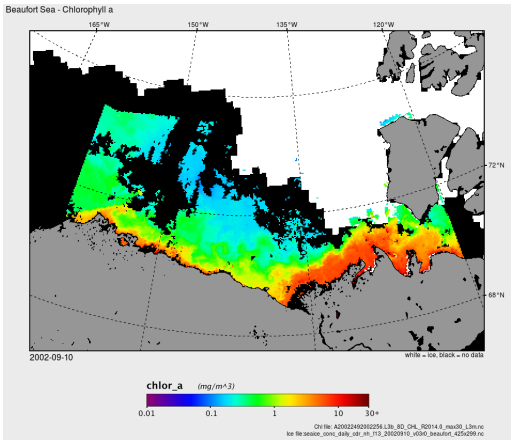

Fig. 1: Plot of chlorophyll a concentrations for one eight-day time step in the data

Due to climate change, we expect that in future years (2016 onwards), features will be more similar to the 2015 data than they will be to earlier data. Thus, evaluating our models on the 2015 data will give us a better sense of how well we will be able to forecast chlorophyll a levels in the future. Various scientists at the Center for Ocean Solutions recommended this approach.

IV. METHODS

We explore various regression techniques to see how chlorophyll a concentrations can be directly predicted from the other features. We experiment with standard regression models and weighted regression models, in order to account for the spatial and temporal variations in the data. Because researchers are more interested in predicting the general areas of biological productivity, rather than the exact chlorophyll a concentrations of any given area, we also briefly explore discretization of the dataset and the use of classification methods.

A. Regression

We experiment with five different regression models: standard linear regression, lasso regression, ridge regression, decision trees, and random forests.

In our linear models, we attempt to construct a hypothesis function, h, that outputs our prediction for each data point $x^{(i)}$ based on a learned set of weights θ. Each individual prediction is given by the result of

$$h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + ... + \theta_n x_n^{(i)}$$

We fit our estimates of θ by minimizing the cost function, $J(\theta)$, which we formulate as an ordinary least squares cost where there are m total points in our training set.

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2$$

Vanilla linear regression on remote sensing data is prone to over-fitting due to large feature spaces and complex nonlinear relationships between the features. Regularized linear regression models using regression trees combined with statistical analysis are increasingly used in ecology. [1] We experimented with several regularized linear regression models.

In the lasso regression model, we regularize the learned weights by their L-1 norm. The L-1 penalty pushes weights for less relevant features close to or equal to zero as the model trains. The loss function to minimize is given by:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2 + \lambda||\theta||_1$$

[1]

Ridge regression also regularizes the learned weights, but uses the L-2 norm for regularization rather than the L-1 norm. Since the regularization term here is squared, weights cannot be pushed to zero. The cost function is:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2 + \lambda||\theta||_2^2$$

[2]

In both models, λ is a parameter that controls the amount of regularization. The weight coefficients learned by all linear regression models can tell us about the relative importance of each feature. We also experimented with Decision Trees, another supervised learning method. Decision Trees learn decision rules from the data features by recursively partitioning the data based on features and selecting the feature partitioning that minimizes the following cost function:

$$J(Q,\theta) = \frac{n_{left}}{N_m}H(Q_{left}(\theta)) + \frac{n_{right}}{N_m}H(Q_{right}(\theta))$$

Where θ represents a partitioning based on certain features, H is an impurity function that measures the goodness of the split, N_m is the number of observations at node m, and n_{right} and n_{left} are the number of observations at the right and left portions of the split.

Since decision trees create complex tree structures, they are very prone to over-fitting to the training data. Thus, we also experimented with Random Forests, an ensemble method. Random Forest Regression uses multiple classifying decision trees on various sub-samples of the dataset and then computes the mean prediction of all the decision trees. It controls over-fitting while improving predictive accuracy by allowing the model to learn the importance of features through the decision tree paradigm. [3]

B. Weighted Regression

Levels of chlorophyll a are influenced by the spatial and temporal aspects of the data; chlorophyll concentrations vary both with location and season. Thus, we experiment with locally weighted versions of our regression models. Locally weighted regression is "a procedure for fitting a regression surface to data through multivariate smoothing: The dependent

[1] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

[2] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

[3] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

variable is smoothed as a function of the independent variables in a moving fashion analogous to how a moving average is computed for a time series." [7] In our case we experimented with both location and time of year as the dependent variables, estimating future levels of chlorophyll a by learning from the independent variables associated with our dependent variable more closely.

We modify our cost functions by multiplying each with a new w_i term, which represents the weight we assign to sample $x^{(i)}$. We model with three weighting schemes: location-based weighting, time-based weighting, and combination weighting. In location-based weighting, we weight each sample by its inverse Euclidean distance from the query point. This means that samples that are closer to the query point will be assigned higher weights than samples that are farther from the query point. Similarly, in our time-based weighting scheme, we weight sample points by the inverse number of days from the query point. In the combination weighting scheme, we combine these two weights so that query points that are closer in both latitude/longitude as well of day of year are given higher weight. We expect that this weighting will improve the performance of our models, since the data varies significantly across time and space.

We utilize the `sklearn` implementations of all models. We evaluate our models by measuring their Mean Squared Error (MSE) and R^2 values. The MSE tells us the average deviation of our predicted values from the actual values – thus, a lower MSE indicates better performance of the model. The R^2 value tells us, roughly, what fraction of the variance in the data our model accounts for. Models that yield higher R^2 values are considered better than models that yield lower values. We did not perform tuning on the hyperparameters in our models (λ values in regularized regression models) – we left them at the default values provided by `sklearn`. While this is a valuable avenue for future work, we chose to focus on the overall performance of each model to try to understand which was overall best suited for the data at this time.

C. Classification

The primary application of interest is in predicting areas of biological productivity. Thus, while it's important that a model is able to predict what the chlorophyll a concentration in a given area is, it's also important that the model is able to distinguish areas of higher productivity from areas of lower productivity. Thus, we briefly experiment with a multi-class classification model for this problem. We discretize the chlorophyll a concentrations into buckets, and then used a Naive Bayes model to make predictions. The Naive Bayes model makes the assumption that all features are conditionally independent given the class, which in this case is the chlorophyll a concentration bucket. Given a feature vector $x^{(i)}$, we compute the posterior probability

$$p(y^{(i)} = c|x^{(i)}) \propto p(x^{(i)}|y^{(i)} = c)p(y^{(i)} = c)$$

for each class c, and then predict the class that results in the highest posterior probability.

V. Results and Discussion

A. Classification

Figures 2 and 3 show the distribution of chlorophyll a concentrations in the training set. Since there is a clear skew in the distribution, we take the log of the chlorophyll a concentration values before discretization. This ensures that the model is not able to achieve high accuracy simply by guessing the most frequent class for each prediction. We then transform the values into ten buckets and run a Naive Bayes classifier on the training set and the test set. We achieve a classification accuracy of 0.310. While this does slightly better than we'd expect from a random model, it does not have satisfactory accuracy. In the interest of time, we choose not to continue with this exploration.

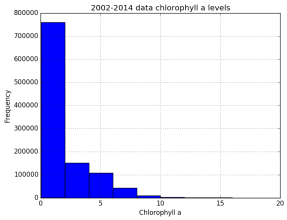

Fig. 2: Train set chlorophyll a concentration frequency

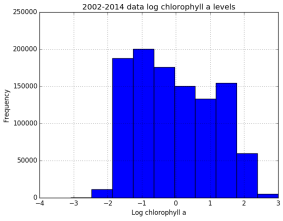

Fig. 3: Train set log chlorophyll a concentration frequency

B. Unweighted Regression

The results of our unweighted regression experiments are summarized in Table I The results of our unweighted regression experiments show that the random forests model performs the best on the test set, achieving the lowest MSE as well as the highest R^2 value. This aligns with the results found by Kotta et al., as boosted regression trees are similar to random forests. [1]

TABLE I: Unweighted regression summary

Model	Train MSE	Train R^2	Test MSE	Test R^2
Linear Regression	3.056	0.339	2.862	0.412
Lasso Regression	3.815	0.175	3.958	0.187
Ridge Regression	3.055	0.339	2.862	2.862
Decision Tree	1.757e-09	0.999	5.474	-0.124
Random Forest	0.047	0.990	2.548	0.477

C. Weighted Regression

Tables II, III, and IV summarize our results from the location-weighted, time-weighted, and combination-weighted models. In our weighted regression models, we present the weighted average MSE over all MSEs yielded by our query points. We define our query location points by generating points at intervals of 5°between the minimum and maximum latitude and longitude. We experimented with a variety of centroid step sizes on a subset of the data, and found that 5°yielded the best results. We generated our time query points by starting with the earliest day in the dataset, and then

selecting further days at 7-day intervals. We test each weighted model only on data points in the test set that fall within a certain radius of the query point (Euclidean distance 1 for location weights, and 7 days for time weights) The MSE and R^2 values presented are weighted averages, weighted by the number of test points associated with each query point. Due to current limitations of `sklearn`, we were unable to experiment with locally weighted lasso regression. In the interest of time, we chose to focus on the other aspects of our project rather than building our own implementation of the model. This is definitely something we'd like to investigate in the future. Our results show that incorporating the time and location weights did improve the performance of our linear regression models, but did not improve the performance of the tree models, and did not improve our best overall baseline performance. The decision tree models consistently overfitted to the train set. This was surprising, given how much the data does vary over time and space. We suspect that the reason why our time weighting did not lead to an improvement is that day of year is a very noisy indicator of temporal patterns in the data. Seasonal patterns vary widely from year to year, so day of year is not the best way of capturing temporal trends. The feature importance coefficients learned by our best model (unweighted random forests) corroborate this.

TABLE II: Location-weighted regression summary

Model	Train MSE	Train R^2	Test MSE	Test R^2
Linear Regression	2.722	0.323	2.612	0.403
Ridge Regression	2.722	0.323	2.612	0.403
Decision Tree	1.6933-09	0.999	6.323	-0.691
Random Forest	0.047	0.988	2.974	0.213

TABLE III: Time-weighted regression summary

Model	Train MSE	Train R^2	Test MSE	Test R^2
Linear Regression	2.891	0.357	2.612	0.403
Ridge Regression	2.891	0.357	2.613	0.403
Decision Tree	1.776e-09	0.999	5.978	-0.389
Random Forest	0.047	0.989	2.995	0.319

TABLE IV: Location and time-weighted regression summary

Model	Train MSE	Train R^2	Test MSE	Test R^2
Linear Regression	2.852	0.278	2.726	0.162
Ridge Regression	2.852	0.278	2.727	0.163
Decision Tree	1.716e-09	0.999	5.842	-0.676
Random Forest	0.048	0.987	2.935	0.218

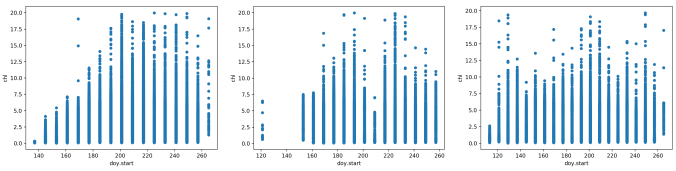

Fig. 4: Chlorophyll levels vs. starting day of year in 2012, 2014, and 2015, showing erratic changes.

The coefficients with the lowest weights are day of year start and day of year end. Figure V show the coefficients learned by the model for each of our features.

TABLE V: Feature importance coefficients learned by unweighted random forest model

Feature	Coefficient	Feature	Coefficient
Latitude	0.421	Max PAR	0.046
Longitude	0.115	Distance to land	0.033
Depth	0.102	Hours of daylight	0.013
Mean PAR	0.068	Sea ice	0.012
Sea surface temperature	0.062	Day of year start	0.005
Cloud cover	0.058	Day of year end	0.004

Future work on this project might include a different representation of the time feature – perhaps by somehow aggregating the different temporal-based features in the dataset (day of year start/end, hours of daylight). However, features such as sea surface temperature and sea ice are rapidly from year to year even at the same day of the year due to climate change. It's therefore useful to make predictions using many of these features discretely.

The coefficients learned by the random forest model indicate that the most important features are the location features – latitude and longitude. Interestingly, however, when we incorporate latitude/longitude based weights into the model, it performs worse on the test set. Perhaps Euclidean distance alone isn't a sufficient amount of information. The model may have been learning different location-based weights on its own that our weights disrupted. However, the incorporation of Euclidean distance weights did improve the performance of our linear regression models, indicating that the weights were valid indicators of chlorophyll a concentrations. Furthermore, depth was another important feature learned by the random forest regression model. Sea depth changes with distance from the coast, and is thus a location-related feature. Figures 5 and 6 illustrate the predictions our location- and time-weighted models made on the 2015 test data, averaged over time.

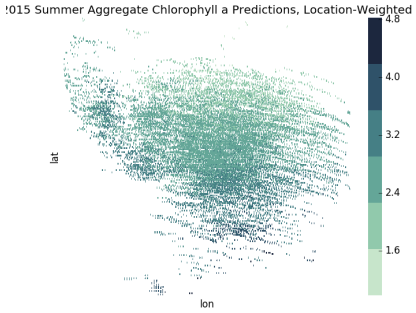

Fig. 5: Location-weighted predictions

D. Seasonal Forecasting

Due to the wide variation of seasonal patterns between years, we investigate the possibility of making predictions within a single season. To do this, we focus on data only from the year 2015. We try training our model on information

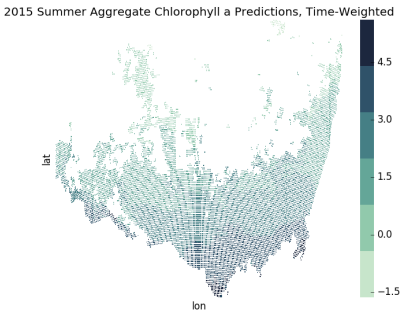

Fig. 6: Time-weighted predictions

gathered from the first part of the year, and then testing on the later part of the year. We split the 2015 data into a train and test set by choosing the first 70 percent of the data as train and the remaining as test. We then run the same regression models on this new subset of the data. Our results are summarized in table VI

TABLE VI: Unweighted regression summary – within 2015

Model	Train MSE	Train R^2	Test MSE	Test R^2
Linear Regression	2.259	0.540	11.006	-1.585
Lasso Regression	2.996	0.390	3.037	0.286
Ridge Regression	2.259	0.540	11.002	-1.585
Decision Tree	9.600e-10	0.999	3.213	0.245
Random Forest	0.041	0.991	2.171	0.490

VI. FUTURE WORK

Our results thus far have shown that while forecasting chlorophyll a concentration is a difficult problem, there are underlying patterns in the remotely sensed and other collected data that relate to chlorophyll a concentrations and biological productivity. The main avenue for future work is through further analysis and work with the different features in the dataset. Our results show that location-based features are certainly informative of biological productivity in areas. Furthermore, while date of year is a noisy indicator of time, other time-based features may have potential to be more predictive when used as a weighting factor. We're interested in seeing how methods such as PCA/ICA can shed more light on the properties of the different features in our dataset. We are also interested in experimenting with this correlation based feature selection algorithm developed at the University of Waikato in New Zealand. It can be applied to discrete classification problems as a preprocessing step before any of our non-weighted regression algorithms and would be interesting to try. [8]

Due to the effects of climate change, we know that the earliest data in our dataset (2002) looks very different from the data in our test set. Furthermore, there is widespread variation in seasonal patterns from year to year. Thus, it may be valuable to invest time into constructing several models, one for each season, or set of days in the year. Perhaps we could experiment with training our model only on more recent data, to fully account for any effects climate change may have. It may also be possible to map "early season data" and "late season data" between various years, accounting for the differences in seasons year to year that are not currently accounted for by day of year. If temperatures continue to fluctuate so much, however, when summer starts may not be the greatest source of temporal fluctuation.

Our results indicate that our best-performing model considered latitude as the most important feature. Visually, we can see that chlorophyll concentrations in this area of the Beaufort Sea vary most dramatically across latitudes. We're interested in seeing how well this model can generalize to other parts of the world. Specifically, there is another dataset collected from the Palau area as part of this project – we're interested in seeing how our model performs there.

ACKNOWLEDGMENT

We would like to thank our mentor, Lisa Wedding, for all her help this quarter. We are also grateful to Andy Stock and everyone on the Catalyst Arctic Project and the Stanford Center for Ocean Solutions for their guidance and assistance.

REFERENCES

[1] J. Kotta, T. Kutser, K. Teeveer, E. Vahtme, and M. Prnoja, "Predicting species cover of marine macrophyte and invertebrate species combining hyperspectral remote sensing, machine learning and regression techniques," *PLOS ONE*, vol. 8, pp. 1–11, 06 2013.

[2] L. Ye and Q. Cai, "Forecasting daily chlorophyll a concentration during the spring phytoplankton bloom period in xiangxi bay of the three-gorges reservoir by means of a recurrent artificial neural network," *Journal of Freshwater Ecology*, vol. 24, no. 4, pp. 609–617, 2009.

[3] E. E. Elattar, J. Goulermas, and Q. H. Wu, "Electric load forecasting based on locally weighted support vector regression," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, pp. 438–447, July 2010.

[4] H. Sun, H. Liu, H. Xiao, R. He, and B. Ran, "Use of local linear regression model for short-term traffic forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1836, pp. 143–150, 2003.

[5] G. Camps-Valls, L. Bruzzone, J. L. Rojo-Alvarez, and F. Melgani, "Robust support vector regression for biophysical variable estimation from remotely sensed images," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, pp. 339–343, July 2006.

[6] B. Haack and A. Rafter, "Regression estimation techniques with remote sensing: a review and case study," *Geocarto International*, vol. 25, no. 1, pp. 71–82, 2010.

[7] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, pp. 596–610, 1988.

[8] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, (San Francisco, CA, USA), pp. 359–366, Morgan Kaufmann Publishers Inc., 2000.