Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230447

Multivariate polynomial regression prediction of primary productivity on U.S. coastal areas

Yabei Zeng

College of letters and Science, University of California, Santa Barbara, CA 93106 United States

yabei_zeng@ucsb.edu

Abstract. This paper analyzes the factors influencing the primary productivity of U.S. coastal areas in 2010. Understanding primary productivity is crucial to our understanding of the ecosystem and the biodiversity it supports. In this analysis, five parameters were taken into consideration, including longitude, latitude, total nitrogen, total phosphorus, and ammonia. Incorporating data from the Nation Coastal Condition Assessment (NCCA), this study mainly focused on the two models, multivariate polynomial regression and random forest, and general performance on the prediction accuracy. Comparing the results of the two models, the multivariate polynomial model does perform slightly better than that of the random forest model. Both of the two models yield good prediction models that can be further applied to the prediction of primary productivity rate for the ecology prediction and biodiversity prediction.

Keywords: Primary Productivity, Multivariate Polynomial Regression, Random Forest, Prediction.

1. Introduction

Primary productivity is the rate at which energy is converted into organic substances by photosynthetic and chemosynthetic producers [1]. Generally, many factors can significantly or slightly affect the rate of conversion. Both nitrogen and phosphorus are the key determining factors in affecting the changes in primary productivity[1]. Though ammonia and geological variations do not significantly contribute to the primary productivity as the prior two do, these two features are interaction terms in better understanding the changes in nitrogen and phosphorus. There are research on terrestrial primary productivity. In “Predicting Gross Primary Productivity in Terrestrial Ecosystems”, the researchers mainly explore terrestrial primary productivity and the diverse ecosystem [2]. In “Vegetation characteristics and primary productivity along an arctic transect: implications for scaling up”, the researchers mainly explore the relationship between vegetation characteristics and primary productivity [3]. This research on the coastal primary productivity explores the primary productivity rate of the coastal areas and the relationship between the primary productivity and the chemical compounds available in the area. Since the original plot of the primary productivity data retrieved from NCCA in 2010 did not have linear relationships in two-dimensional spaces based on the five features, non-linear models, multivariate polynomial regression, and random forest, are incorporated to predict the primary productivity rate. The prediction of primary productivity helps understand the ecosystems of the area

DOI: 10.54254/2755-2721/6/20230447

and can further be applied to learn about biodiversity. Primary productivity can be important and related to food availability, so it can be further applied to predict species abundance.

2. Methodology

2.1. Dataset semantics and structure

In the dataset from the Nation Coastal Condition Assessment, there are in total of 1091 observational units after data cleaning [4]. In the cleaned dataset, there are 7 categorical variables and 6 numeric variables listed below with specific descriptions (As shown in Table 1).

Table 1. Data description.

Name	Variable Description	Type
SITE_ID	The identification code where the data is collected	Categorical
WTBDY_NM	The country where the data is collected	Categorical
COUNTRY	The country where the data is collected	Categorical
STATE	The state where the data is collected	Categorical
PROVINCE	The province where the data is collected	Categorical
DATE	The date of data collected	Categorical
NCCR_REG	The region of the data collected	Categorical
LAT	The longitude of the site	Numeric
LON	The latitude of the site	Numeric
Ammonia	The total ammonia detected	Numeric
Productivity	The total chlorophyll A detected	Numeric
Total Nitrogen	The total nitrogen detected	Numeric
Total Phosphorus	The total phosphorus detected	Numeric

2.2. Methods

2.2.1. PCA analysis and dimension reduction. PCA projects the data into a low-dimensional subspace to achieve dimensionality reduction. For example, a two-dimensional dataset is reduced by projecting the points into a line, and each sample of the dataset can be represented by a single value without two values. A three-dimensional dataset can be reduced to two dimensions, which means mapping the variables into a plane [5]. Noticing that the dataset has 5 features while only containing 1092 observations, which is too small for 5 features, using PCA dimension reduction helps to project the variables into a new subspace that allows us to have fewer dimensions to operate with [6]. By using PCA dimension reduction, it is possible to "construct a d x k–dimensional transformation matrix W that allows us to map a sample vector x onto a new k–dimensional feature subspace that has fewer dimensions than the original d–dimensional feature space" [6]. In this process, PCA prevents the original dimensions from decreasing the accuracy of the prediction.

2.2.2. Multivariate polynomial regression. The primary benefit of polynomial regression is that it may be approximated by adding higher terms of x to the real points until a reasonable fit is achieved. Any

DOI: 10.54254/2755-2721/6/20230447

function may be approximated by polynomials in segments, hence polynomial regression can handle a large class of nonlinear problems and has a significant position in regression analysis. Consequently, regardless of the relationship between the dependent variable and other independent variables, polynomial regression can always be used for analysis in common practical situations [7]. The multivariate polynomial regression, which furthers the calculation based on polynomial regression by adding more variables to the matrix. It is approximately the same as the matrix below but with second or third variables added, increasing degrees.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i \ (i = 1, 2, \ldots, n) \tag{1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ 1 & x_3 & x_3^2 & \cdots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

Figure 1. The matrix of polynomial regression [8].

Given that the plot presents a non-linear trend, "polynomial regression can be used when the response is not linear." [8]. In this situation, when the dataset contains more than one predicting variables, polynomial regression can be applied and "computed on multiple regressors as multivariate polynomial regression" [9]. Multivariate polynomial allows us to have the lowest cost value and avoid overfitting.

2.2.3. Random forest. Random forest here works to provide a comparable index for the results of the multivariate polynomial regression. Random forest consists of a large number of decision trees that work as ensembles. Every single tree in the random forest "spit out with a class prediction and the class with the most votes become our model's prediction" [10].

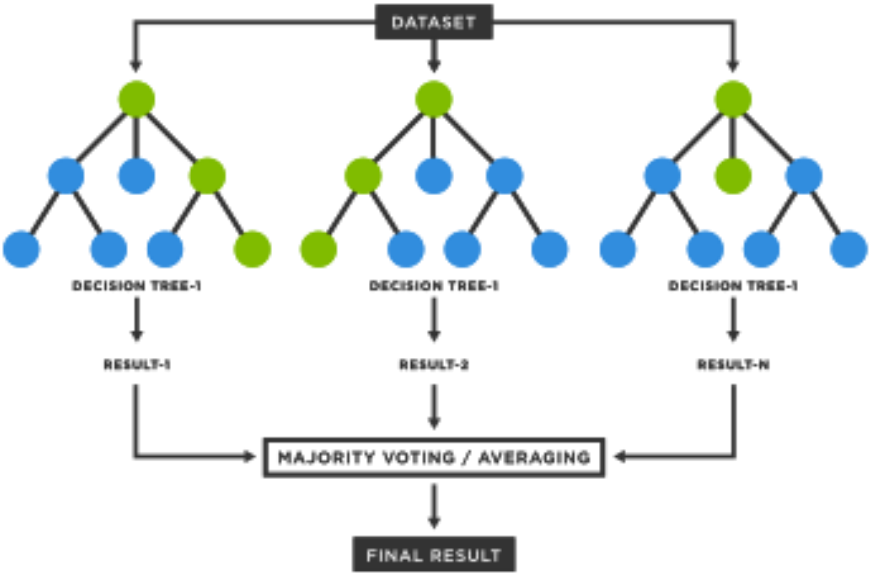

Figure 2. Random forest structure [11].

DOI: 10.54254/2755-2721/6/20230447

3. Analysis

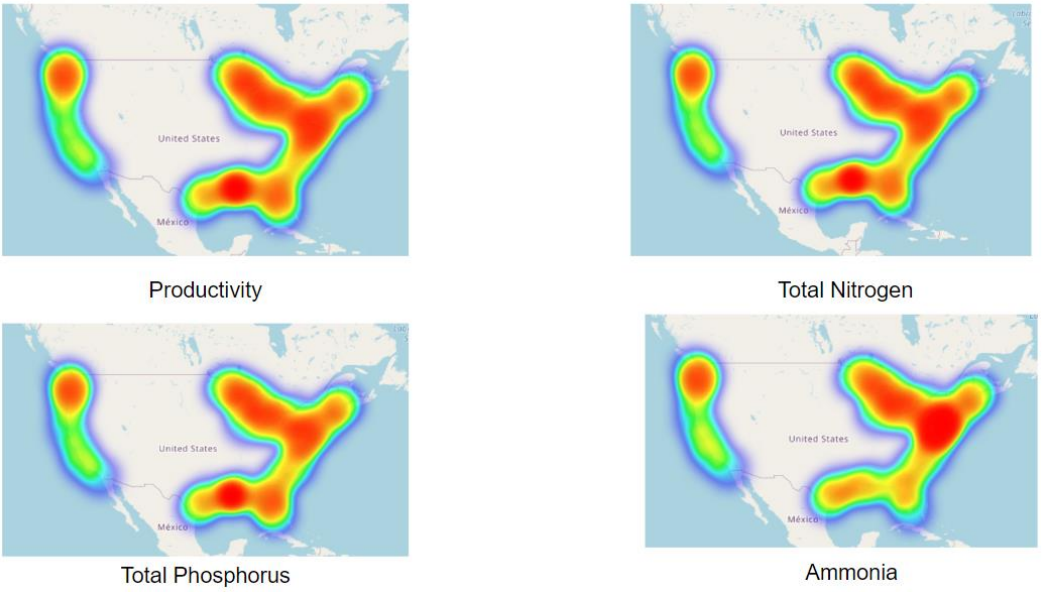

Figure 3. The visual plot of Primary Productivity, Total Nitrogen, Total Phosphorus, and Ammonia based on latitude and longitude.

In the four plots, the warm color represents a higher value, whereas the cold color represents a lower value. It is clear that primary productivity increases as the rest three increase and vice versa. Additionally, for nitrogen, phosphorus, and ammonia, their values also vary due to the geological variations. Though it is possible to tell the vague relationship between primary productivity and the remaining features, the detailed relationship is still unknown.

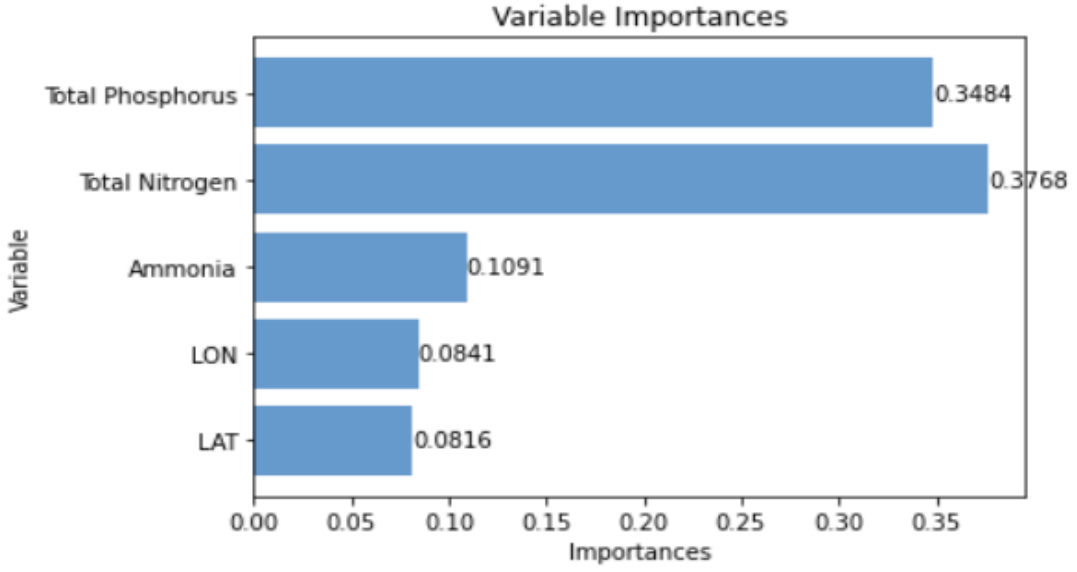

Figure 4. The variables importance.

In this plot, it can show that both total nitrogen and total phosphorus contribute to primary productivity the most, whereas ammonia, longitude, and latitude contribute relatively small to primary productivity but are still non-negligible. As the plot showed total phosphorus contributes 0.3484, total nitrogen contributes 0.3768, ammonia contributes 0.1091, longitude contributes 0.0841, and latitude contributes 0.0816.

DOI: 10.54254/2755-2721/6/20230447

3.1. Multivariate polynomial

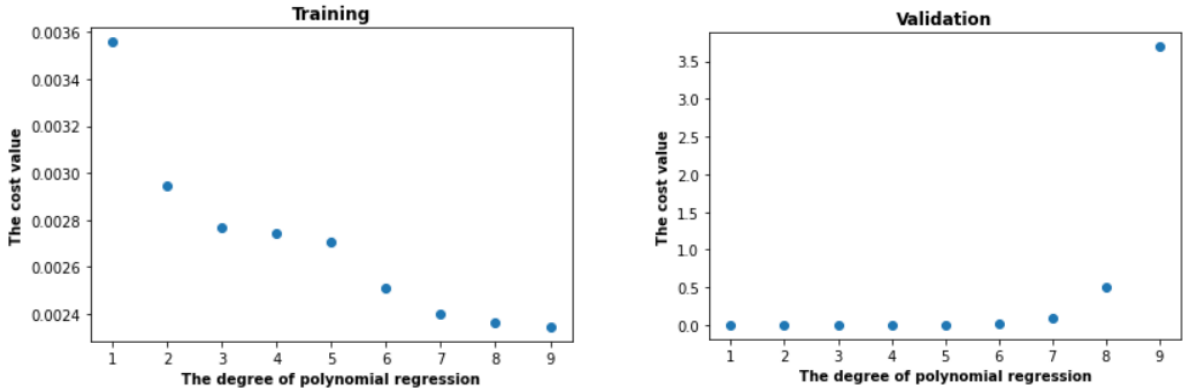

Figure 5. The cost value of training dataset(left) and validation dataset(right).

In polynomial regression, when the degrees of the variables increase, the prediction line would be more likely to fit the training dataset by connecting each point. Therefore, in the cost value plot of the training dataset, when the degree is higher, the cost value is correspondingly decreasing. As the plot shows, when the degree is 1, the cost value is higher than 0.00334 and when the degree is 9, the cost value is lower than 0,024. However, the higher degrees, the more likely it will be overfitting. From the validation plot, where the models of the training dataset are applied to the validation dataset, the cost value first decreases then increases. Therefore, degree 2 where the cost value achieved the lowest in the validation is chosen in the final model of multivariate polynomial regression.

```
R2 score:  0.7278956058632621
RMSE:  0.0028331924150925567
   Actual values  Predicted Values
0        0.00298          0.002084
1        0.00839          0.009664
2        0.01502          0.011258
3        0.00030          0.001316
4        0.01002          0.008980
```

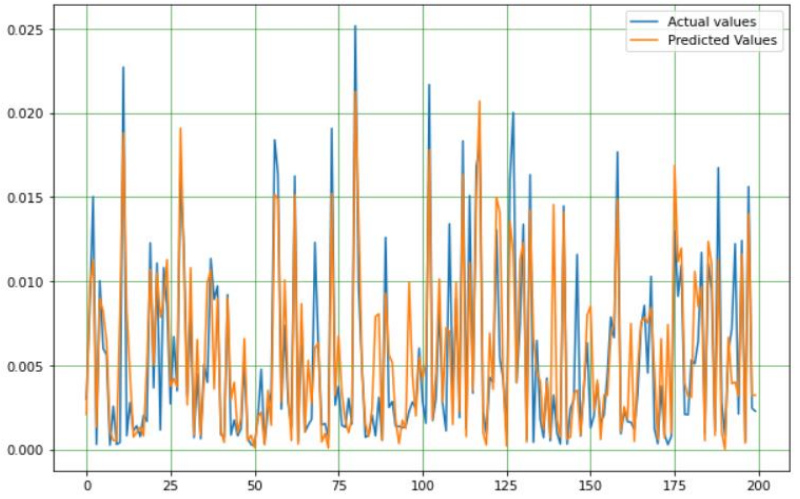

Figure 6. Multivariate polynomial regression: actual values and predicted value comparison.

From the plot above, the r2_score, the total variance explained by the model over the total variance, is near 0.7279. The mean squared error, or the cost value, is around 0.0028. With the R2 score being closer to 1 and in the range between 0 and 1 and with mean squared error being small, the model of multivariate polynomial regression has yielded expected predictions that are close to the actual values.

DOI: 10.54254/2755-2721/6/20230447

3.2. Comparison with random forest

```
R2 score:  0.7241153918110199
RMSE:  0.0028528046167134837
   Actual values  Predicted Values
0        0.00298          0.002526
1        0.00839          0.010178
2        0.01502          0.011933
3        0.00030          0.001217
4        0.01002          0.008382
```

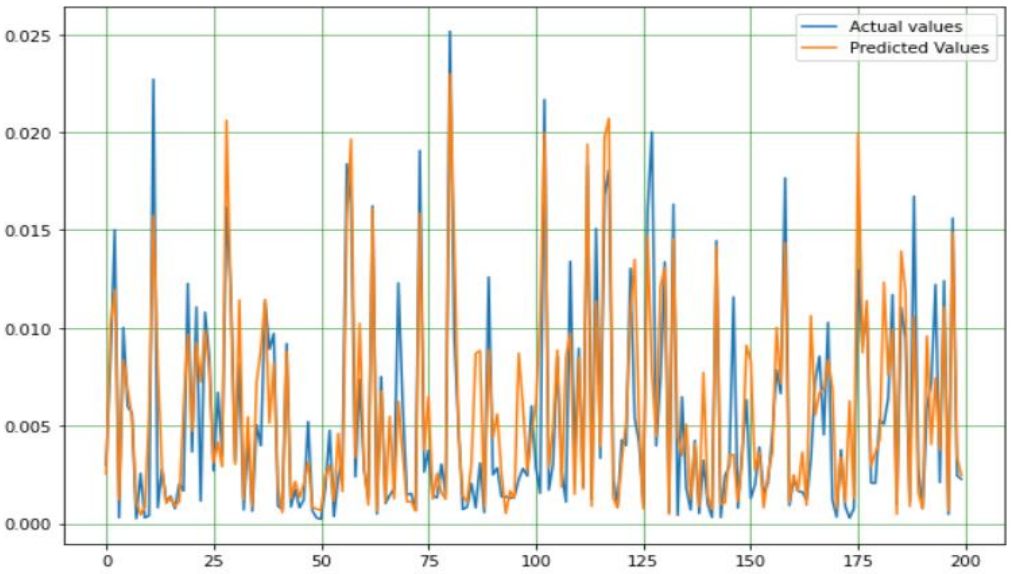

Figure 7. Random Forest: actual values and predicted value comparison.

In the random forest model, the R2 score is near 0.7241 and the mean squared error is around 0.00285. The R2 score is slightly lower than that of the multivariate polynomial regression, and the mean squared error is slightly higher than that of the multivariate polynomial regression. Both models have produced good predictions that are approximately close to the actual values.

4. Discussion

The prediction may be less accurate for primary productivity in other countries. Noticing that variables' contributions are important to the prediction models, other chemical compounds that are not noticeable available in the United States but have importance in other countries' coastal areas may be important to the new prediction model. However, based on the algorithm and the methods for the prediction of coastal primary productivity in other countries, it is possible to yield good results by reevaluating the variables based on their importance.

5. Conclusion

From all the plots above, all of the numeric variables, total nitrogen, total phosphorus, ammonia, latitude, and longitude, have been related to the rate of primary productivity, with nitrogen and phosphorus contributing more significantly. With producing interacting terms regarding the five parameters, both multivariate polynomial regression and random forest have produced a good prediction model of primary productivity. Comparing the two models, the performance of multivariate polynomial regression is slightly better than that of the random forest. Overall, the results showed are sufficient to predict the value approximately close to the actual ones. However, there is still ways to enhance the accuracy of the prediction. As the model and the datasets suggest, only the importance variables are taken into consideration, but there may be other significant contributor, such as pollutants, are not included in the dataset. Therefore, in the further research, the model can be bettered by adding more available variables and enhance the dataset. Also, as the primary productivity is crucial to the ecosystem

DOI: 10.54254/2755-2721/6/20230447

and species abundance. Therefore, in the further research, the prediction models and the results can be applied in the prediction of species abundance of the areas.

Acknowledgment

First and foremost, I would like to express my heartfelt appreciation to my supervisor, a respected, responsible, and resourceful scholar who has provided me with invaluable guidance throughout the composition of my thesis. I could not have finished my thesis without his illuminating instruction, outstanding kindness, and patience. His sharp and ardent academic observation enlightens me not only in this thesis but also in my future research. I'd like to thank Mrs. Liu for her warmth and assistance. I would also like to thank all of my professors for assisting me in developing fundamental and vital academic skills. My heartfelt gratitude also goes to my supervisor, who engaged in this study with excellent cooperation.

References

[1] The Editors of Encyclopaedia Britannica, 2022, Primary Productivity, Retrieved September 2nd, https://www.britannica.com/science/primary-productivity

[2] Mathew Williams, Edward B. 1997. Rastetter, David N. Fernandes, Michael L. Goulden, Gaius R. Shaver, Loretta C. Johnson, Predicting Gross Primary Productivity in Terrestrial Ecosystems, Ecological Applications, 7(3), 882-894.

[3] Mathew Williams, Edward B. 1999. Rastetter, Vegetation Characteristics and Primary Productivity Along an Arctic Transect: Implications for Scaling-Up, Journal of Ecology, 87(5), 885-898.

[4] EPA 2022, Data from the National Aquatic Resource Surveys, Retrieved September 2nd, https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys

[5] Kiran Parte 2020, Dimensionality Reduction Principal Component Analysis, Retrieved September 2nd, https://medium.com/analytics-vidhya/dimensionality-reduction-principal-component-analysis-d1402b58feb1

[6] Lorraine Li 2019, Principal Component Analysis for Dimensionality Reduction, Retrieved September 2nd, https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad

[7] Tamas Ujhelyi 2021, Polynomial Regression in Python Using Scikit-Learn, Retrieved September 2nd, https://data36.com/polynomial-regression-python-scikit-learn/

[8] Wikipedia 2022, Polynomial Regression, Retrieved September 2nd, https://en.wikipedia.org/wiki/Polynomial_regression

[9] Priyanka Sinha, 2013. Multivariate Polynomial Regression in Data Mining: Methodology, Problems and Solutions, International Journal of Scientific and Engineering Research, 4(12), 962-965.

[10] Tony Yiu 2019, Understanding Random Forest, Retrieved September 2nd, https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[11] TIBCO 2022, What is a Random Forest, Retrieved September 2nd, https://www.tibco.com/reference-center/what-is-a-random-forest