Linear Regression in SLR and Coastal Economy

2024 Data Science Competition

Team: Statistical Anomalies
Minh Dao Nguyen
advisor: Jobin Varughese

April 5, 2024

1 Executive Summary

This paper use Linear Regression model to explore the relationship between the sea level data and coastal states economy. The states of focus are East coast states. The data collected are from 2 well established sources. This report will go through the data gathering process, the methodology, the modeling and analysis, visualization, and the conclusion. Supplement materials that completed this reported are included in the folder of the of this paper.

2 Problem Statement

This paper explores the relationship between the sea level data from Coperinus Climate Change Service and the coastal economy.

3 Datasets

Table 1: Copernicus Climate Change Service — DUACS delayed-time altimeter gridded maps of sea surface heights (SSH) and derived variables over the global Ocean.

Variable	**Description**
adt	Sea surface height above geoid [m]
sla	Sea surface height above sea level [m]
ugos	Surface geostrophic eastward sea water velocity [m/s]
ugosa	Surface geostrophic eastward sea water velocity assuming sea level for geoid [m/s]
vgos	Surface geostrophic northward sea water velocity [m/s]
vgosa	Surface geostrophic northward sea water velocity assuming sea level for geoid [m/s]

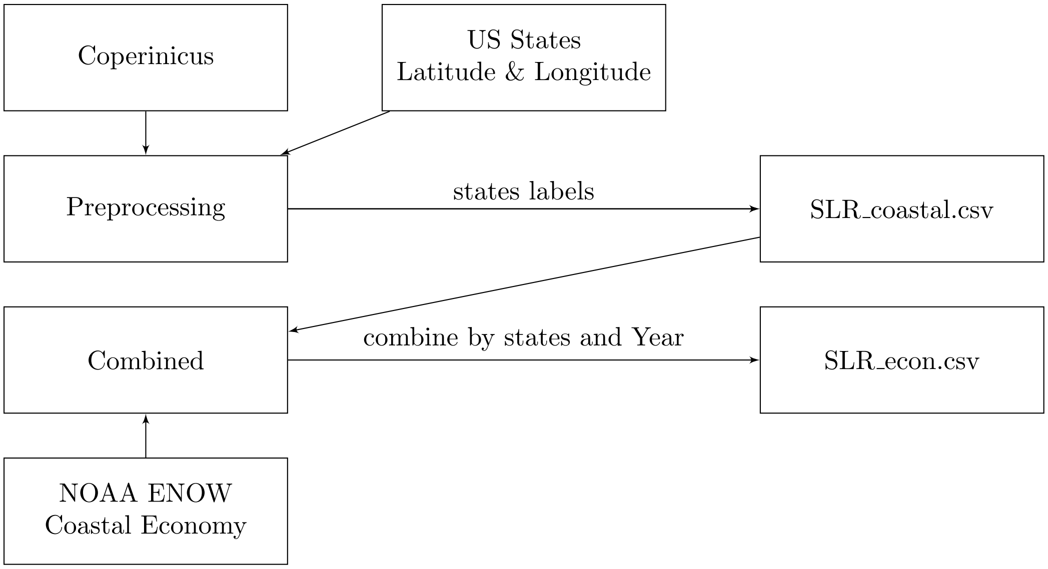
Table 2: NOAA ENOW Explorer - Coastal State Economy data for six sectors dependent on the ocean and Great Lakes: Tourism and Recreation, Ocean Economy, Living Resources, Offshore Mineral Resources, Ship and Boat Building

Variable	**Description**
geoName	Name of the geographic region
year	Year of observation
sector	Sector of economic activity
establishments	Number of business establishments
employment	Number of employed individuals
wages	Wages of employed individuals
gdp	Gross Domestic Product

Table 3: Variables describing latitude and longitude of US states

Variable	**Description**
State	State abbreviation
State Name	Full name of the state
Lat_min	Minimum latitude
Lat_max	Maximum latitude
Long_min	Minimum longitude
Long_max	Maximum longitude

4 Data Exploration



5 Methodology

The main method use to analysis the relationship between the sea level data and the coastal economy is linear regression model. This method is very simple model and light weight for computation, which work well with fitting less variables. Square-root transformation is also used in order to improve the data and the variance.
Other advance Machine Learning method has been considered, such as Neural Network or Supported Vector Machine (SVM). However, the data is not big enough to uztilize those advance model, especially Neural Network. This could cause overfitting in bigger model. However, we did use some decision trees and boosting method to test our their validity. The reason we chose these models because they tend to do better with weak-learner, which is what we believe that this dataset has due to its size and variables. These analysis aren't not the main focus of the paper, but rather just a extra exploratation. The data will be run through these models in Python using Sklearn libraries.

6 Modeling and Analysis

6.1 Linear Regression

6.1.1 Model Fitting

This exploration paper will mostly focus on Linear Regerssion. The goal is to fit a Linear model and minimize the residuals sum of square. We will also examine which variables or predictors are significant in predicting wages and gdp. Multiple linear regression is used to fit the model, where the formula is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Where:

- Dependent variable Y: wages
- Predictor $X_1, X_2, \ldots, X_n$: adt, sla, ugos, ugosa, vgos, vgosa, year, establishments, employment
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_n$: the coefficients predictors X's.

- ϵ: the error term

We composed the model in R using the `lm` function as follows:

```
model <- lm(wages ~ adt + sla + ugos + ugosa + vgos + vgosa + year + establishments + employment, data =
```

To obtain a summary of the model, we used the `summary` function:

```
summary(model)
```

This provided us with detailed information about the fitted linear regression model, including coefficients, standard errors, t-values, p-values, and goodness-of-fit statistics.

Table 4: Coefficients

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(> $\vert t \vert$)**
(Intercept)	-2.017e+11	1.032e+11	-1.954	0.0508 .
adt	-7.882e+09	1.454e+09	-5.423	6.38e-08 ***
sla	3.828e+09	4.186e+09	0.914	0.3605
ugos	-1.735e+10	6.920e+09	-2.508	0.0122 *
ugosa	1.985e+10	9.439e+09	2.103	0.0355 *
vgos	-8.940e+09	7.151e+09	-1.250	0.2114
vgosa	1.579e+10	8.525e+09	1.852	0.0641 .
year	1.001e+08	5.127e+07	1.952	0.0511 .
establishments	1.419e+05	1.609e+04	8.815	$< 2e - 16$ ***
employment	5.057e+04	1.080e+03	46.847	$< 2e - 16$ ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.061e+10 on 2714 degrees of freedom
Multiple R-squared: 0.9622, Adjusted R-squared: 0.962
F-statistic: 7667 on 9 and 2714 DF, p-value: $< 2.2e - 16$

The result of fitting a linear regression model with all the variables show which of the coefficients have significant impact to the wages. In order to reduce the residuals error of the model, we adopt square-root transformation for the wages variable. The transformation help stabilizing the variance and reduce the influence of outliers. Since different states has different economy system and policy, square-root transformation helps normalize the data.

6.1.2 Square-root Transformation

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 45370
Multiple R-squared: 0.7836, Adjusted R-squared: 0.7829
F-statistic: 1092 on 9 and 2714 DF, p-value: $< 2.2e - 16$

The standard errors has reduced a lot compare to the last model.

6.1.3 Model Analysis

We fitted a linear regression model to the data and obtained the following results: The summary statistics of the linear regression model are as follows:

- The predicted variables that are statistically significant n predicting the wages are sla, year, establishments, employment. Their p-values are less than 0.05.

Table 5: Coefficients

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
(Intercept)	-7.851e+06	4.414e+05	-17.787	$< 2e-16$ ***
adt	1.085e+02	6.216e+03	0.017	0.9861
sla	-7.655e+04	1.790e+04	-4.276	1.97e-05 ***
ugos	2.950e+04	2.959e+04	0.997	0.3189
ugosa	7.301e+03	4.037e+04	0.181	0.8565
vgos	-3.779e+04	3.058e+04	-1.236	0.2167
vgosa	4.400e+04	3.646e+04	1.207	0.2276
year	3.920e+03	2.193e+02	17.878	$< 2e-16$ ***
establishments	1.636e-01	6.883e-02	2.377	0.0175 *
employment	8.102e-02	4.616e-03	17.550	$< 2e-16$ ***

- The predicted variables that are NOT statistically significant in predicting the wages are adt, ugos, ugosa, vgos, vgosa. Their p-values are greater than 0.05.

- Residual Standard Error: 45370 on 2714 degrees of freedom
  The residual standard error is really big, which shows that the predicted or fitted values are not close to the actual values.

- R-squared 0.7836
  The (R^2) implies that 78.36% of the variability in the wages is explained by the predictor variables in the model.

- F-statistic: 1092 and p-value: $< 2.2 \times 10^{-16}$
  p-value < 0.05. The F-statistic test shows that the model is statistically significant, with $\alpha = 0.05$

6.1.4 Model Conclusion

The results implies that the linear regression model fits the data well. The R-square (R^2) has the value of 0.7836, the F-statistic is statistically significant ($p < 2.2 \times 10^{-16}$). The predictors is significantly related to the dependent variable. However, looking at the residuals vs fitted value, the plot is not pattern less. This means that assumptions of the regression model is not being met. This could mean non-constant variance, non-linear relationship, or outliers (which is not likely due to square-root transformation). These issues need to be investigated in order to assure that the model is correct. Therefore, it cannot be concluded that the model is good.

6.2 Other Models

We tried some other Machine Learning models in Python to test theirs ability to fit and predict the data. These models were trained with a very short range of parameters, so they wouldn't be as optimal as they could be.
Random Forest: mse - $3.959304071666519 \times 10^{19}$
Gradient Boosting: mse - $3.3061833507993113 \times 10^{19}$
AdaBoost: mse - $4.750173341792919 \times 10^{19}$
Decision Tree: mse - $4.421627256385497 \times 10^{19}$

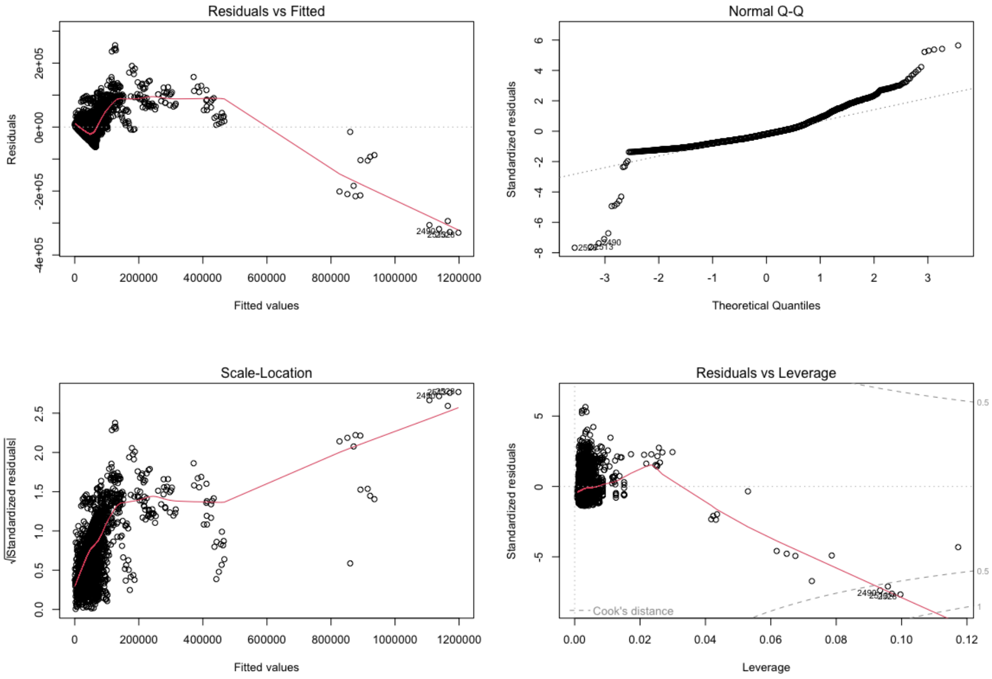


Figure 1: Diagnostic Plots

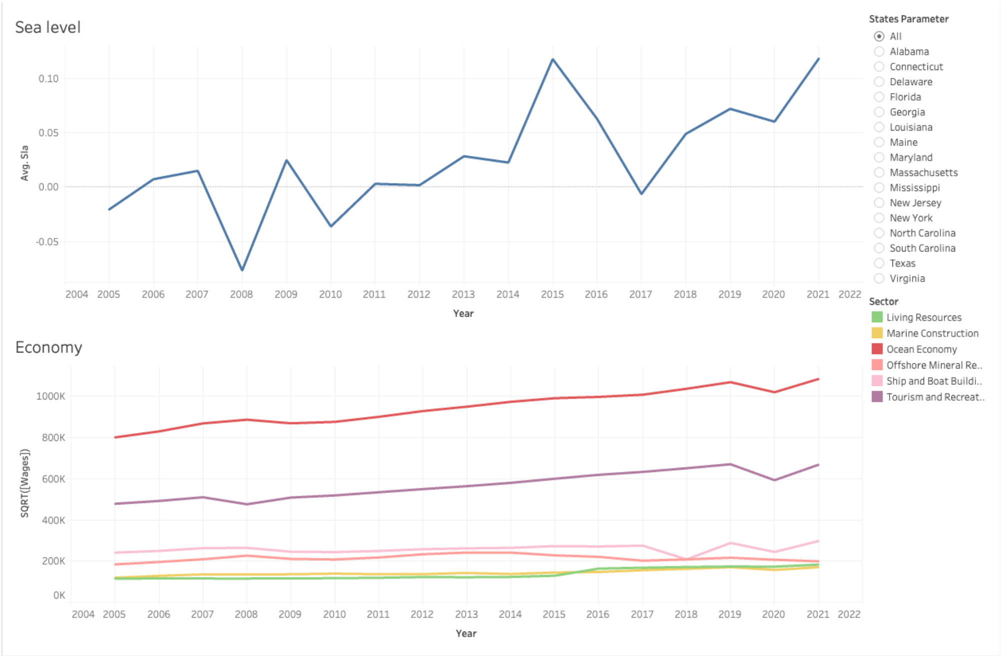


Figure 2: Sea Level and Coastal Economy Sectors

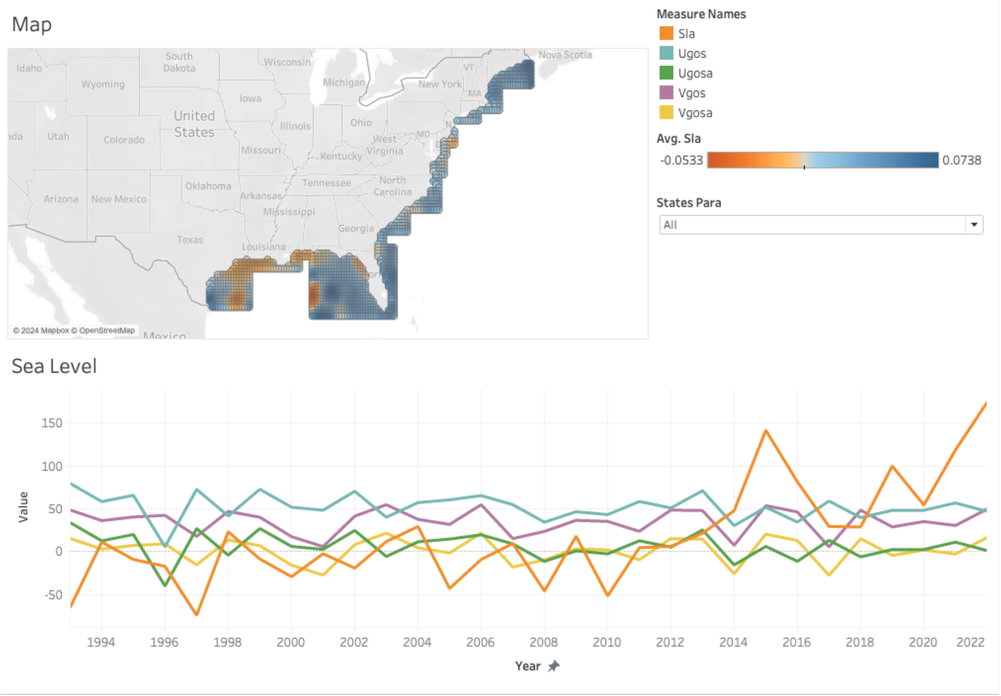


Figure 3: Data Map view of Sea Level

7 Visualization

8 Conclusions and Recommendations

The sea level data set from Copernicus data does somewhat contribute to the predicting the coastal economic impact wage of coastal states. However, it does not fully explain the reason for the coastal economy changes. The sea level has a negative effect on the wages of the coastal economy based on the linear regression model. This could lead to further analysis on the cause of this effect.

Overall conclusion of this paper is that there is not a strong enough relationship or evidence to use sea level rise data from Copernicus to predict coastal states economy.