TAMU Sea Data Challenge Report

By: Daniel Chen and Kenhao Chin

Link to Project: https://github.com/xxilytoo/TAMUSEADataChallenge

Executive Summary

The tide gauge data from Dangendorf et al. (2023) has important hidden patterns within waiting to be uncovered so that future rising sea levels can be predicted. We approached the problem to predict future sea levels by using vector autoregression model. The model was trained on variables we analyzed through correlation visualization.

Problem Statement

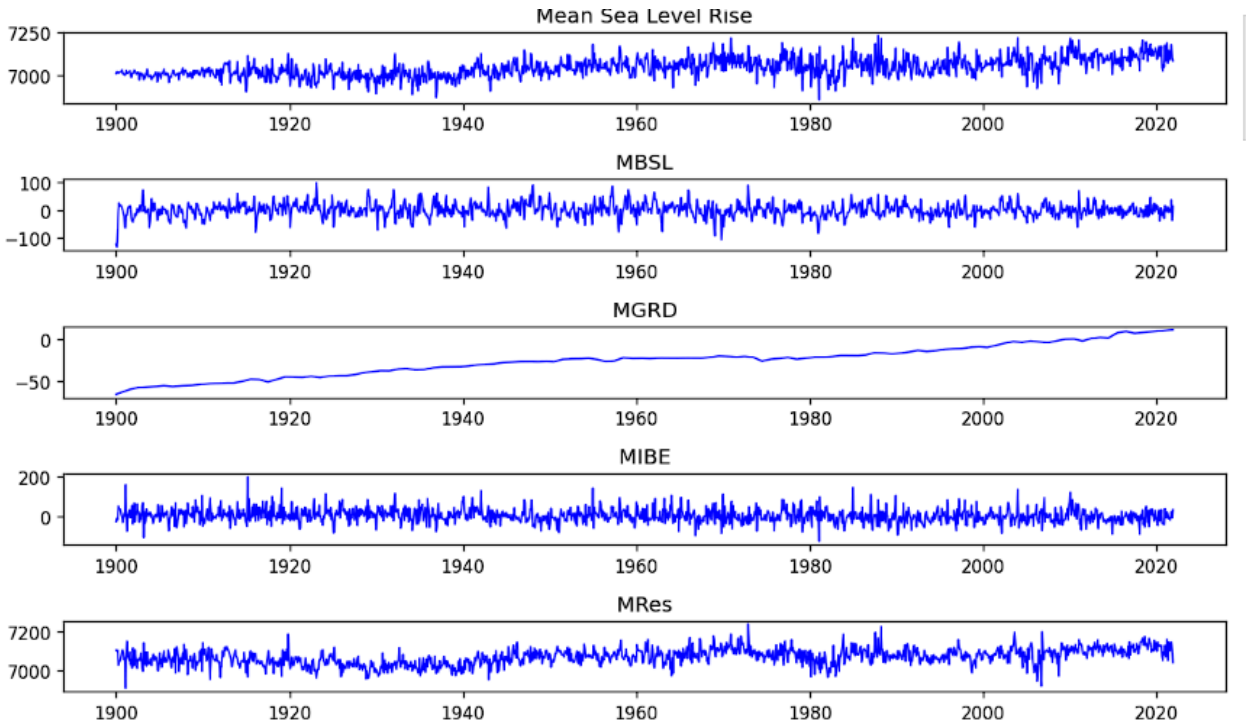
The specific research problem we decided to tackle was predicting the SLR impact. We approached this by creating machine learning models in order to predict the rise of sea level and then analyzed how these predicted sea levels would impact life around the area.

Datasets

We chose to use the Tide Gauge data from Dangendorf et al. to train our machine learning models. To process the data, we first started by reading the data from the matlab file into csv files so that we could more easily access the data using python. After loading the data into csv files, we proceeded to plot the data in order to visualize it.

Data Exploration

To explore the data, we initially created a plot of sea level versus time. This served as a baseline and showed how sea level changed relative to the time. We also plotted the other 4 variables (MBSL, MGRD, MIBE, and MRes) in respect to time. These graphs allowed us to gain a better understanding of how each variable correlates with time. These plots offer good insight but comparing 2 variables in isolation may not hold meaning while using multiple variables might yield different results.



Methodology

We chose a vector autoregression model because we wanted a forecasting model that could predict variables based on time and the relationship between each other. The training for the vector autoregression model used the location marked “fb3t5ddw3” after the geohashing in data preprocessing.

The first test we used is the Augmented Dickey-Fuller test to determine whether or not the data was stationary. If the data was not stationary, we would adjust the data by taking the differenced value. The data was adjusted until the Augmented Dickey-Fuller test resulted in a p-value less than 0.05, meaning it is stationary.

The next test used for the vector autoregression model is the Granger Causality Test. This test is used to determine the minimum number of lags the model must be trained on for two different variables to have causation on each other. The lowest lag value that has a p-value less than 0.05 should be used.

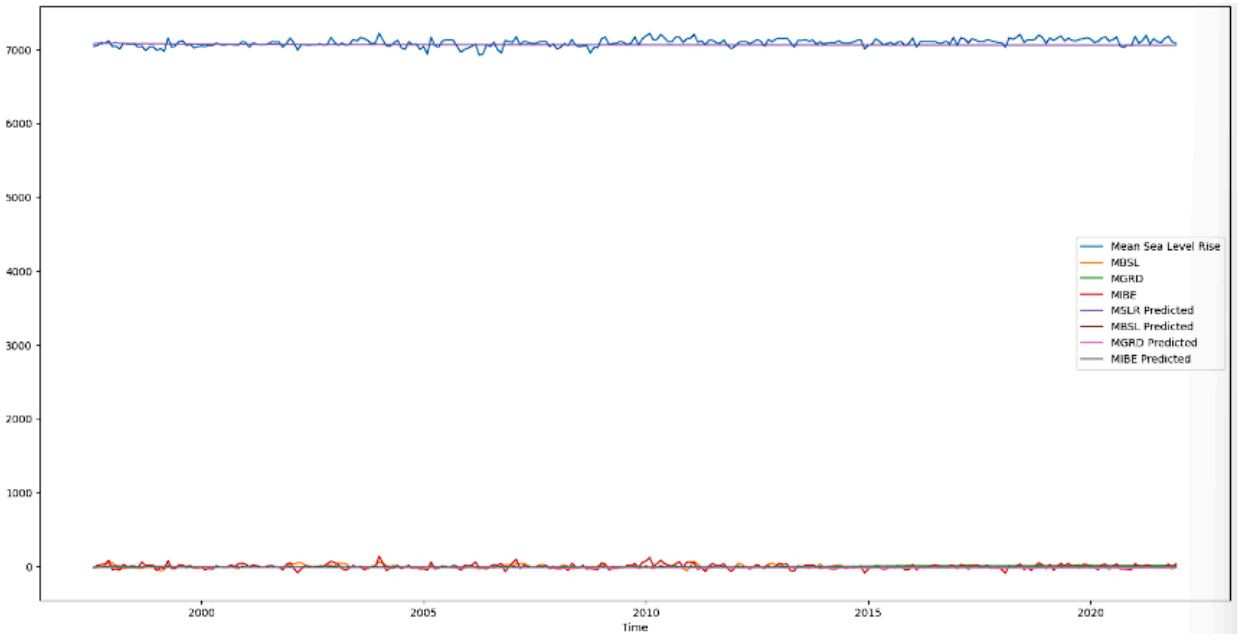
Finally, we find the optimal lag length to use by testing for the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Final Prediction Error (FPE), and Hanna-Quinn Information Criterion (HQIC). The Akaike Information Criterions measures how well the model fit while giving penalty for the number of parameters. The BIC is almost similar to the AIC, but it places a stronger penalty. FPE is the estimation of the model’s error variance and HQIC is also similar to AIC but favors simpler models compared to AIC. The lag number marked with the most asterisks (*) in each row means that lag number is the most optimal to train the model on.

Modeling & Analysis

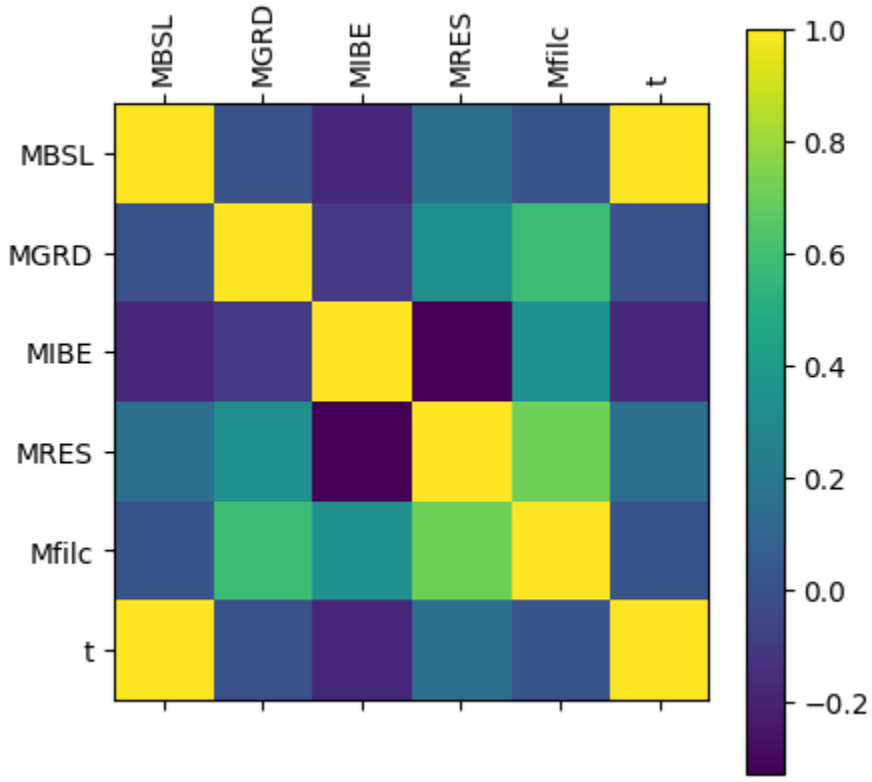
VAR Order Selection (* highlights the minimums)

	AIC	BIC	FPE	HQIC
0	16.95	16.97	2.295e+07	16.96
1	14.64	14.72	2.274e+06	14.67
2	14.30	14.46	1.627e+06	14.36
3	14.21	14.44	1.480e+06	14.29
4	14.08	14.38*	1.300e+06	14.19
5	14.05	14.41	1.260e+06	14.19
6	14.04	14.47	1.246e+06	14.20
7	14.00	14.51	1.206e+06	14.19
8	13.96	14.54	1.152e+06	14.18
9	13.93	14.58	1.120e+06	14.17*
10	13.92	14.64	1.109e+06	14.19
11	13.91	14.70	1.100e+06	14.21
12	13.91	14.77	1.102e+06	14.24
13	13.82*	14.75	1.008e+06*	14.17
14	13.83	14.83	1.016e+06	14.21
15	13.83	14.91	1.020e+06	14.24
16	13.84	14.98	1.021e+06	14.27
17	13.84	15.05	1.028e+06	14.30
18	13.84	15.12	1.027e+06	14.33
19	13.84	15.20	1.030e+06	14.35
20	13.86	15.28	1.044e+06	14.39

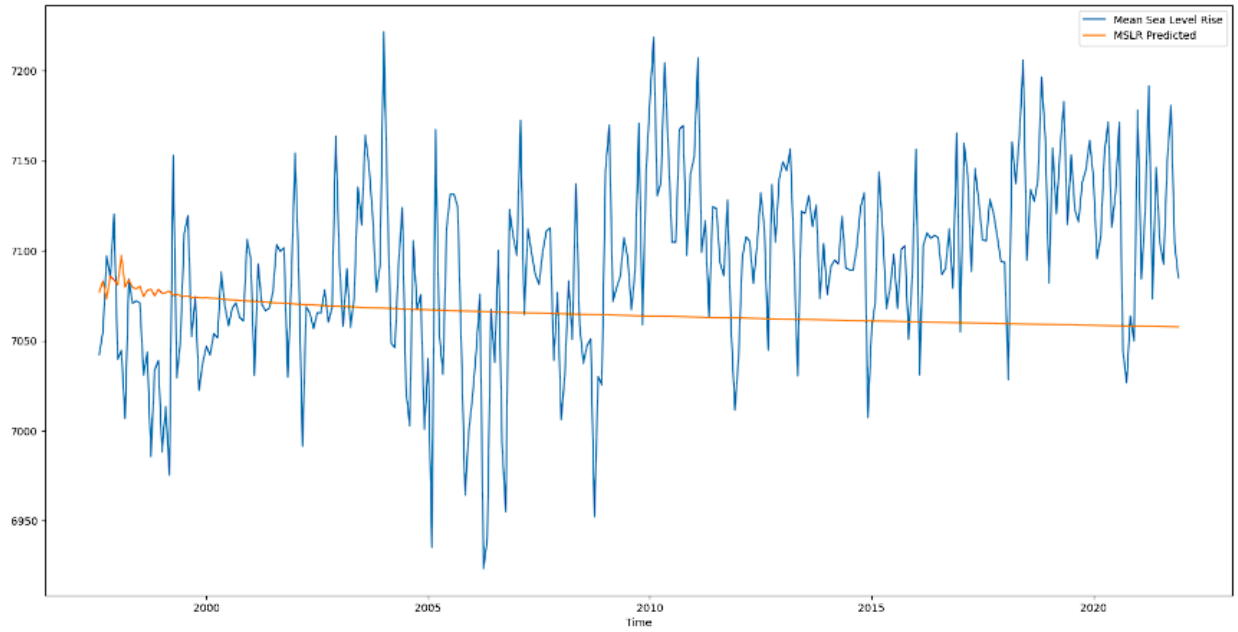
For the vector autoregression model, the data was split into 80% train and 20 % testing. A lag of 13 was chosen to train the model since it was the most ideal lag from the AIC, BIC, FPE, and HQIC. In the middle of training the model, we dropped the Sterodynamic residual factor because we could not get the data to become stationary since it takes more lags than can be tried. The main statistic we used to test the error was root mean square error. We only tested the error for both Mean Sea Level Rise (MSLR) and Barystatid GRD (MGRD). The mean for MSLR was around 7091.85 and the root mean square error was 59.83. On the other hand, the mean for MGRD was around 0.101 and the root mean square error was 17.92. The chart below shows the prediction for each variable and its original values over the course of time.



Visualization & Interpretation



From the correlation visual, it appears that MBSL is very correlated with MBSL, a one to one positive correlation. It is also apparent that Mfilc is very correlated with MGRD, MIBE, and MRES. MRES exhibits some correlation with MBSL and MGRD displays negative correlation with MIBE. This visualization allows us to make surface level analysis which can be further expanded on by using a machine learning model to forecast future sealevels.



This visual is the zoomed in version of the predicted MSLR compared to the true values. As seen in the chart, the predicted values do not follow the volatility of the true values. We realized that the model was only sticking around the mean value of the various data points. After thorough analysis, we realized that our time values had been left out of the model training since it did not fit the date-time format that python required. Due to the tight schedule, we could not find a way to correctly incorporate it into the model.

Conclusions & Recommendations

Even though our model's MSLR root mean squared error was only around 59.83 which is within reason compared to the mean of 7091.85, we find that our forecasting vector autoregression model does not follow the volatility of the true values. In the future, we will correctly implement the time values into the date-time format so that the model can be trained more accurately. After improving the model's accuracy, we plan to train other model for the different locations and create a web application for researchers to forecast the sea levels at different locations.

Resources Used:

- https://core.ac.uk/download/pdf/145047426.pdf
- https://phdinds-aim.github.io/time_series_handbook/03_VectorAutoregressiveModels/03_VectorAutoregressiveMethods.html
- https://doi.org/10.5281/zenodo.7749568