Clustering and fitting of Nitrous Oxide Emissions

Srinivas Vegisetti

Introduction

- World bank dataset containing the yearly emissions of nitrous oxide was utilized.
- Dataset was filtered to remove rows not representing the countries and columns with no data (completely missing).
- The data was brought to a proper structure with some data and data frame manipulation.
- Data was read from, pre-processed, normalized, created logistics formula and err_ranges with the help of user defined functions.
- The data contains 217 countries and other areas along with 13 attributes representing years
- There were no missing values in the data detected with the usual isna() method.
- The missing values were present with period symbols instead of zero or NaN. They were detected by checking the data type of columns and converted to integer. If that failed, the period symbols were replaced with 0.

Clustering

- I compared the recent years of 2018 and 2019 with 8-18 years older (1990 and 2000).
- Based on the silhouette score, 2 clusters were the best for both old and new years.
- Old years best silhouette score: 0.9396
- New years best silhouette score: 0.9484
- For old years, 214 samples were together in one cluster while 3 samples were together in the other cluster. Those 3 samples are China, India and United States.
- For new years, 213 samples were together in one cluster while 4 samples were together in the other cluster. Those 3 samples are Brazil, China, India and United States.
- There is some similarity between the results of old and new years i.e. China, India and United States are always together in a different cluster from the rest of the countries.
- Since China, India and United States are different from most of the samples, they may be considered outliers or anomaly points.
- Their emissions are much larger than the remaining countries.

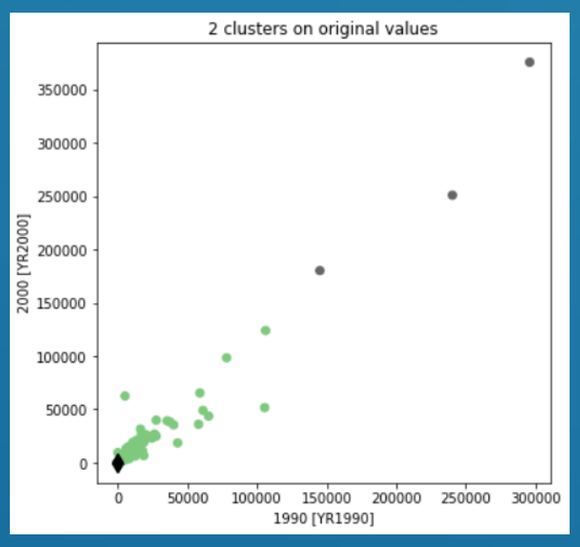

Figure1: Old years clustering

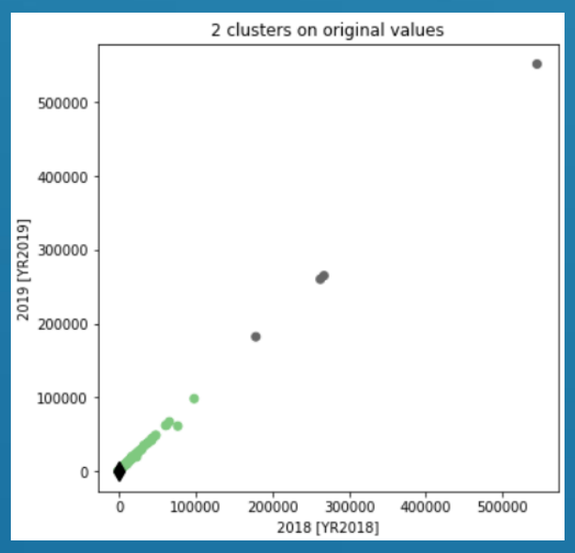

Figure 2: New years clustering

Logistics Curve Fitting

- As a continuation to the clustering part, the logistics function was fit to two countries from different clusters to facilitate comparison in the traits of the clusters.
- India and Zimbabwe were chosen.
- India was the representative from the minority cluster
- Zimbabwe was the representative from the majority cluster.
- Created 2 data frames, one for each country.
- Data transformation was done to create 2 columns
- First represented the year and the second represented the amount of emissions in that year.

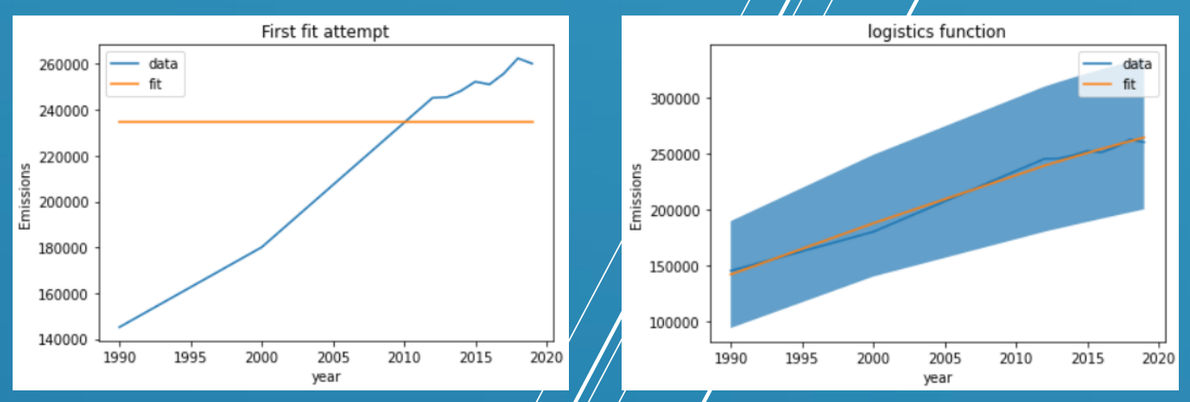

Figure 3: India Logistic Curve Fitting

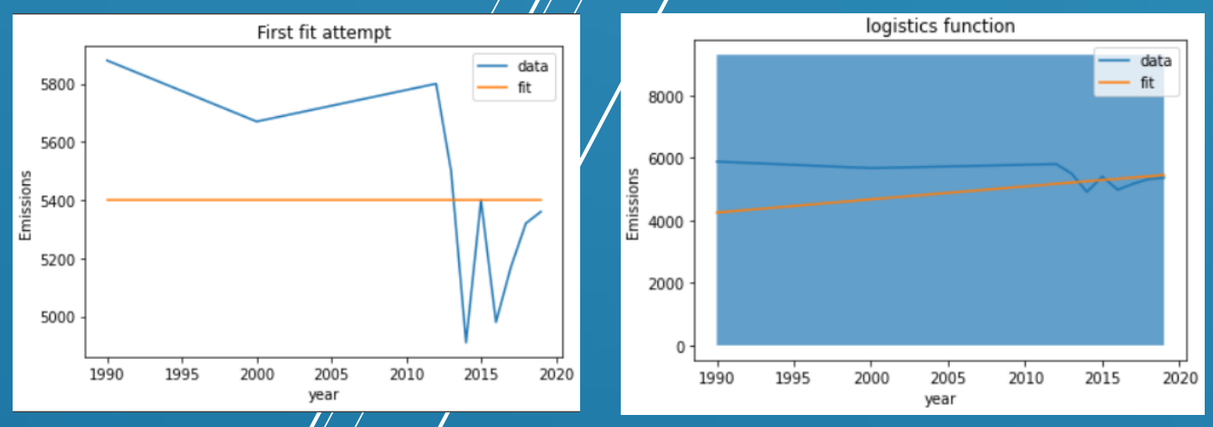

Figure 4: Zimbabwe Logistic Curve Fitting

India's Forecasted Emissions

Year	Lower bound	Upper bound
2030	227583.91	361091.24
2040	247513.45	374341.12
2050	262681.58	381618.43

Zimbabwe's Forecasted Emissions

Year	Lower bound	Upper bound
2030	0.63	9295.337
2040	1.09	9295.333
2050	0.24	9295.317

Curve Fitting Results

India:

- The data exhibited an increasing linear trend in the graph.
- The first fit was almost a constant horizontal line which did not consider any variation in the data.
- The final fit was extremely close to the actual data.
- Except for 2 points the fit overlapped with the data line.
- The error range obtained was quite broad.

Zimbabwe:

- The data exhibited an erratic trend towards the end in the graph.
- The first fit was almost a constant horizontal line which did not consider any variation in the data.
- The final fit was not so close to the actual data.
- The final fit did not overlap very well with the data line.
- The error range obtained was much broader when compared to the India's fitted curve.