Clustering and Curve Fitting

Name: Chimaroke Amaike
ID: 21085699
Repo: https://github.com/daUnborn/cluster-and-fitting

Introduction

This study employs a clustering algorithm utilizing the K-Means technique from the scikit-learn library to analyze patterns in the percentage of land area for agricultural land. Ultimately, the analysis is to see how agricultural land area has increased, reduced or remained stagnant in different countries of the world over a period of 40 years i.e., from 1980 to 2020.. The study further applies the technique of curve fitting to the data in order to determine the underlying agricultural land trend, which is subsequently utilized for forecasting future agricultural land projections.

However, a quick look at the chat on the right shows the trend of agricultural land from 1961 till 2020. There is an apparent upward trend but this report would show relationship between countries

Methods

Data Collection: The data that was used for this analysis was collected from the world bank data bank and can be accessed from this link - https://api.worldbank.org/v2/en/indicator/AG.LND.AGRI.ZS?downloadformat=excel

Data Preprocessing: The Pandas library was utilized to extract and clean data from the climate change dataset. The process involved creating a function to handle the extraction and preprocessing of datasets. The data in the dataframe was not normalized, however, normalization was applied while plotting the clustering results. Data integrity was analyzed to ensure suitability for the task at hand, null entries were dropped, and the shape, datatype, and statistical relationship of the data were inspected.

Clustering: The K-Means centroid-based clustering model was employed. This method assumes that the data points are scattered. The number of clusters was determined using the Sum of Squared Errors (SSE) elbow method.

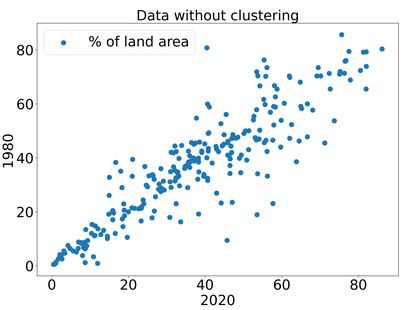
Fitting: A polynomial model was constructed and fed into the curve_fit() function of the Scipy library. Additionally, the err_ranges script was utilized to generate the lower and upper bounds of the fitting data.

Plotting: Data visualization was performed using the Pandas and Matplotlib libraries. A reusable function was created to plot the clustered results with the specified parameters.

Data Analysis

Descriptive Statistics: A description of the data from the clustering can be seen here. This is derived using the describe function. The description of the whole data set was ignored

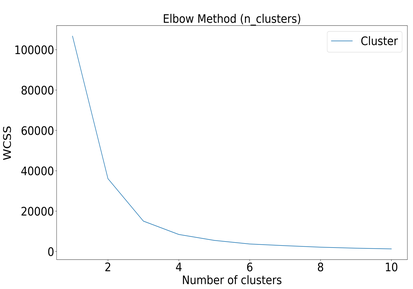
Statistics	1980	2020
count	254	257
mean	37.5799869	37.2093664
std	20.535046	20.2593189
min	0.44230769	0.53846154
25%	21.8514786	21.3333333
50%	37.3095248	38.0667036
75%	53.2608198	48.5469305
max	86.1672952	85.6389987

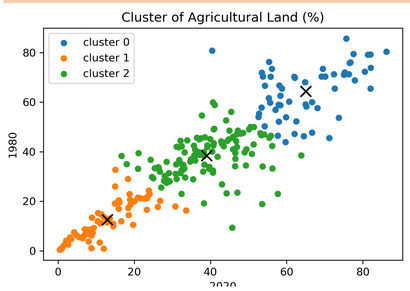


Clustering: First, a scatter plot of the dataset considered was plotted as shown on the left. This gives an insight of what the data looks like before the clustering was done.

We do not know what number of clusters we should adopt. Using the SSE, the clusters were determined by plotting the SSE on a graph. This is also show.

Once the SSE was determined, the clustering was done to show how agricultural land % has faired over the period of 40 years.

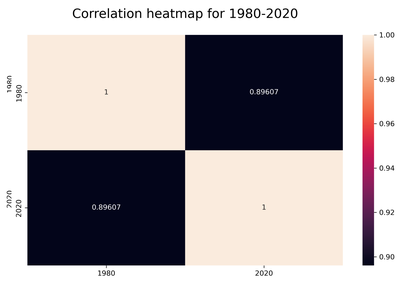




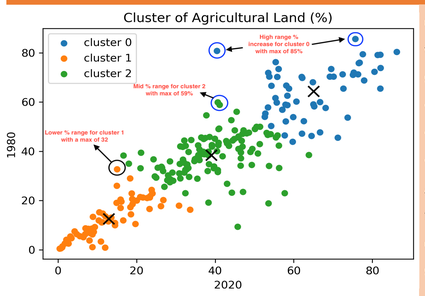
Data Analysis

Correlation:

While analyzing the data, I check for the correlation between the years in review was done. From the plot, it is obvious that the 2 years selected has a strong correlation. As the agricultural land (% of land area) inreases for 1980, it also increases for 2020.



Observation - Clustering

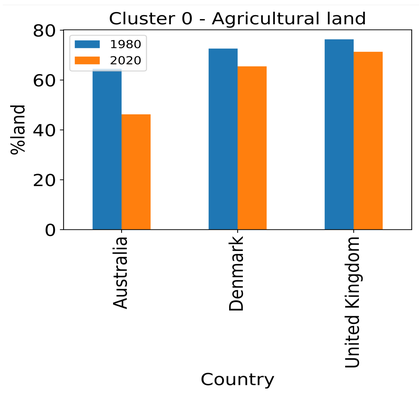


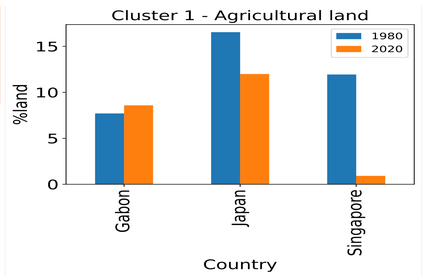
Clustering: From the clustering, the following max values were identified for each of the clusters as shown in the figure;

- Cluster 2 = 59.93
- Cluster 1 = 32.80
- Cluster 0 = 85.64

Also, with a **Silhouette score of 0.5**, we can deduce that the data in the cluster is far from being overlapped. This shows that the data is well separated to a very high degree.

Taking a closer look at selected countries , the following where observed;





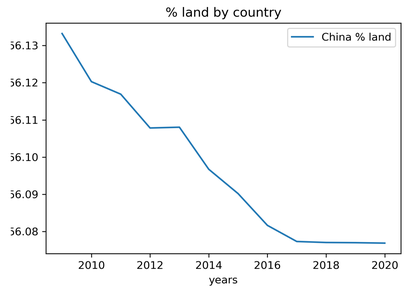
The following were observed on the data;

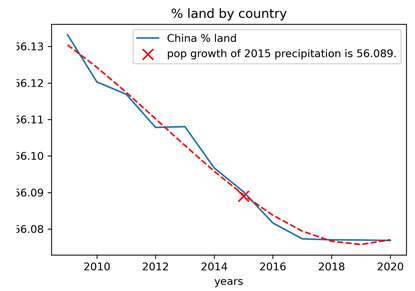
Denmark, Australia, and the United Kingdom are all developed countries with advanced agricultural sectors. They have a mix of modern and traditional farming methods and have diversified agricultural products. They also have a significant amount of arable land that is used for crop production. As agricultural advancement increases, the agricultural land as a percentage of the land increases. However, urbanization and industrialization has ensured that the percentage of agricultural land reduced over the years with Australia being affected the most. Also farms get consolidated. As farms become larger and more mechanized, they may require less land to produce the same amount of food.

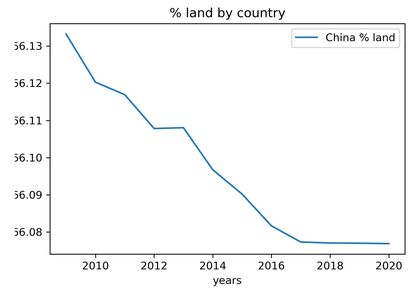
For countries in cluster 1, all three countries have a significant portion of their population living in coastal areas, which has led to urbanization and land reclamation, resulting in a reduction of land area. The limited land area in these countries has led to land scarcity and high land prices, which has led to an increase in land reclamation and urbanization. Japan and Singapore are both located in the Pacific Ring of Fire, which makes them vulnerable to earthquakes, tsunamis and volcanic eruptions. These natural hazards have limited the amount of land available for development.

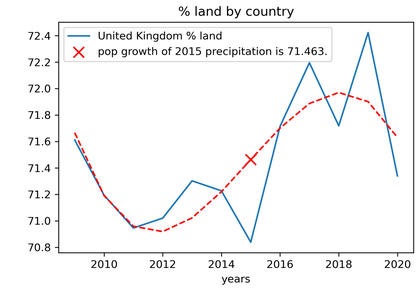
Observation - Fitting

Several countries were sample to see a trend of their agricultural land over a period of 10 years. It was observed that most countries have a decline in their agricultural land as seen in the figures below. You can also observe by the side the **curve_fit** for the trend lines using a polynomial model. On the fit line, there is also a prediction of what the % of land should have been for year 2015

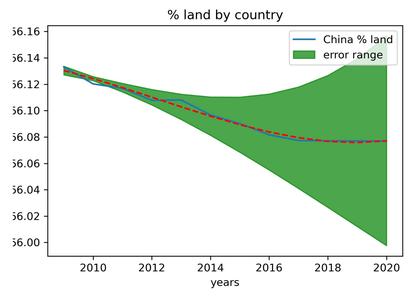


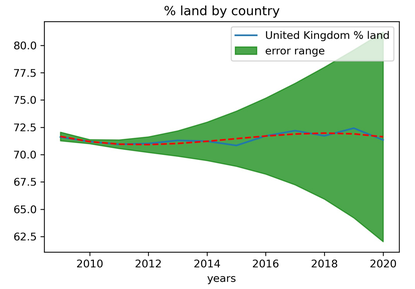






A final look at the error ranges shows that the error margin increases as the agricultural land increase. I was not able to deduce why based on the data given at the time of writing this report





Further Reading

Further work on this is to do the following;

Work on the error ranges to find out why the margin becomes bigger towards the end of the chart

Carry out figure predictions of what land area is used for agriculture as a percentage of total land area

Attempt to use more curve_fit models like expontials and see the kind of result we can get

Carry out a detailed research on the individual countries on why there are declines in their agricultural lands

Further Reading

1.Lebanon Population Decline

2.https://www.fao.org/sustainability/news/detail/en/c/1274219/

3.https://worldpopulationreview.com/country-rankings/countries-by-density