Clustering Analysis of Agricultural Land Use (% of Land Area) in Countries Worldwide

Assignment 3: Clustering and fitting

Faisal Zulfiqar 22010044

Github Link: https://github.com/Faisal-Zulfiqar786/Applied-Data-Science-Assignment-3.git

Abstract

This study groups countries based on their agricultural land use between 1981 and 2021 using cluster analysis. The study utilizes three clustering algorithms, namely K-means clustering, hierarchical clustering, and DBSCAN clustering, to categories countries according to their agricultural land use. The outcomes demonstrate that K-means clustering yielded the best results. The results were also illustrated using parallel coordinate plots and box plots. Researchers and policymakers could use the findings of this study to better comprehend the trends and patterns of agricultural land usage in various countries and areas.

Introduction

Agriculture plays a significant role in a country's economic and social development. Agriculture land area as a percentage of land area is one of the essential indicators of a country's economic and social development. A country with a high percentage of agriculture land area indicates its dependency on agriculture, and a low percentage indicates the development of other industries and services.

Agricultural land use refers to the use of land for agricultural activities such as crop production, livestock rearing, and forestry. It is an essential component of food production and contributes significantly to the economies of many countries. However, agricultural land use is also a significant contributor to climate change. Deforestation, soil degradation, greenhouse gas emissions, and other environmental issues are linked to agricultural activities.

In this work, we will analyze the percentage of agriculture land area of different countries over the years 1981 to 2021. We will perform different clustering techniques to group the countries based on their percentage of agriculture land area and visualize the results. The purpose of this analysis is to understand the trends in the percentage of agriculture land area in different countries and how the countries are clustered based on this indicator.

Dataset

The Agricultural Land Use dataset from the World Bank provides information on the percentage of land used for agriculture (% of land area) for various countries around the world. The data is available from 1960 to 2020 and is obtained from the World Development Indicators (WDI) dataset. The data is compiled from various sources, including national statistical offices, United Nations agencies, and World Bank staff.

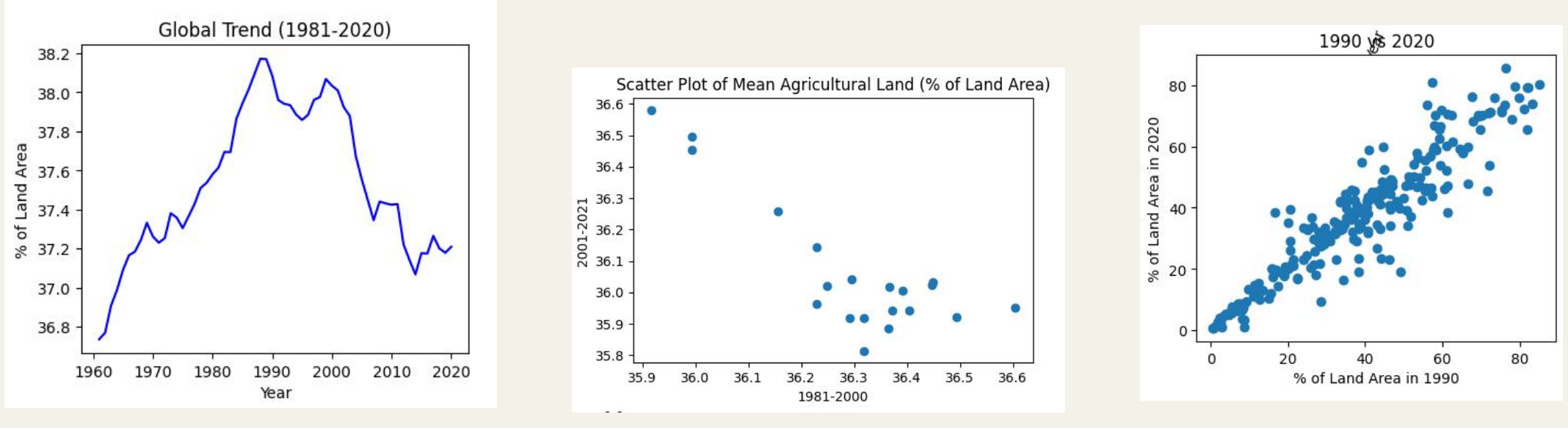


Figure 1: Dataset Exploration

The dataset contains 264 rows, one for each country in the dataset, and 62 columns, one for each year from 1960 to 2021. The mean value for the entire dataset is 36.086%, with a standard deviation of 20.304%. The minimum value is 0.000%, and the maximum value is 94.765%.

The data is preprocessed by filling in missing values with zeros. The data is then scaled using the StandardScaler method from the scikit-learn library to ensure that the different features are on the same scale.

Clustering

Clustering is a common unsupervised machine learning technique used to group similar data points into clusters based on their features or characteristics. The goal of clustering is to identify patterns and structures in the data without any prior knowledge of the groups or categories that the data points belong to.

There are various types of clustering algorithms, but three commonly used ones are K-means, hierarchical clustering, and DBSCAN.

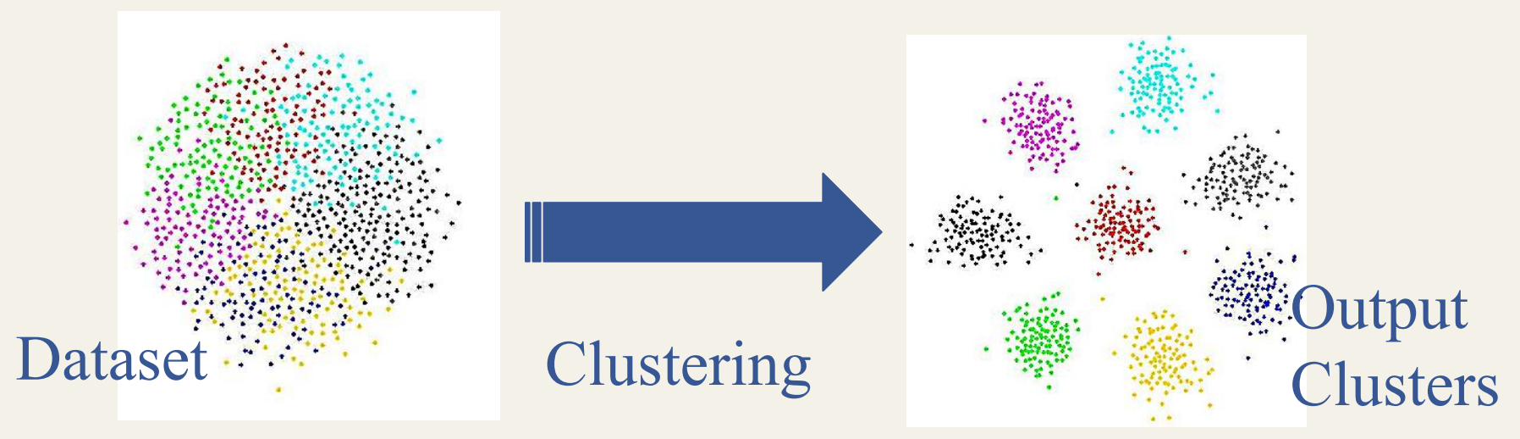


Figure 2: Clustering

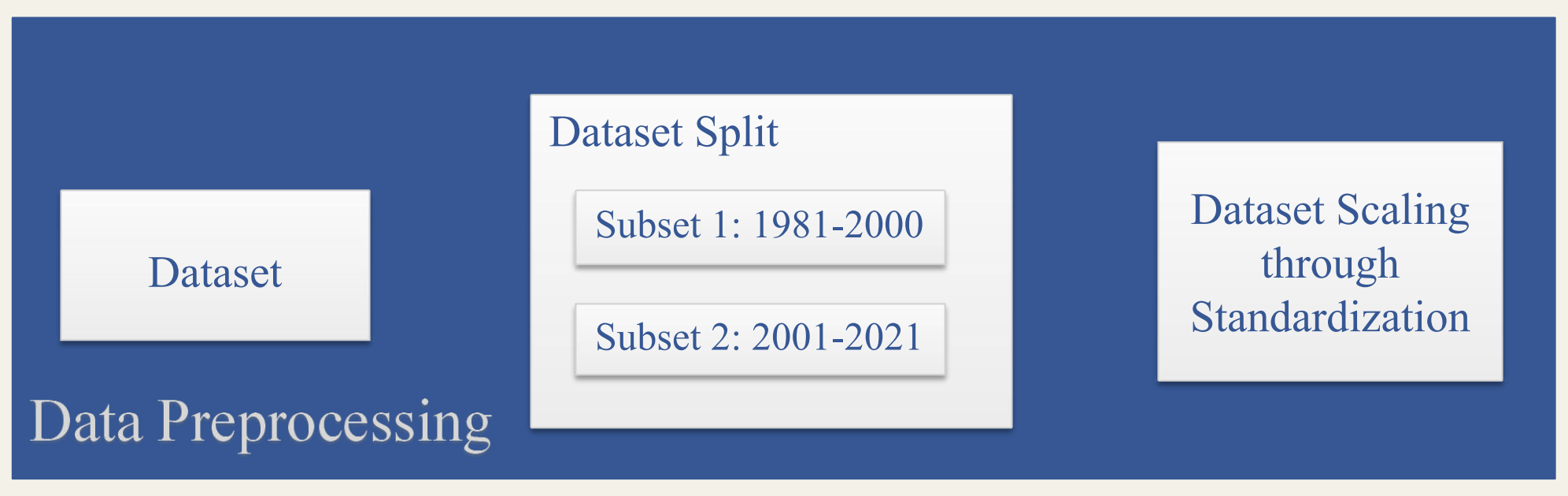
The strengths and weaknesses of each clustering algorithm are as follows:

- K-means is easy to implement and can work well with large datasets, but requires the number of clusters to be specified in advance, and may converge to suboptimal solutions depending on the initialization of the centroids.
- Hierarchical clustering can handle various shapes of clusters and does not require the number of clusters to be specified in advance, but may be computationally expensive and sensitive to noise and outliers.
- DBSCAN can detect clusters of arbitrary shapes and does not require the number of clusters to be specified in advance, but may be sensitive to the choice of hyperparameters such as eps and min_samples, and may not work well with datasets that have varying densities.

Methodology

This work involves clustering of agricultural land use data from the World Bank dataset using three clustering algorithms, namely K-means, hierarchical clustering, and DBSCAN.

The data was subset into two time periods, 1981-2000 and 2001-2021, and then scaled using standardization.


Dataset

Dataset Split

Subset 1: 1981-2000

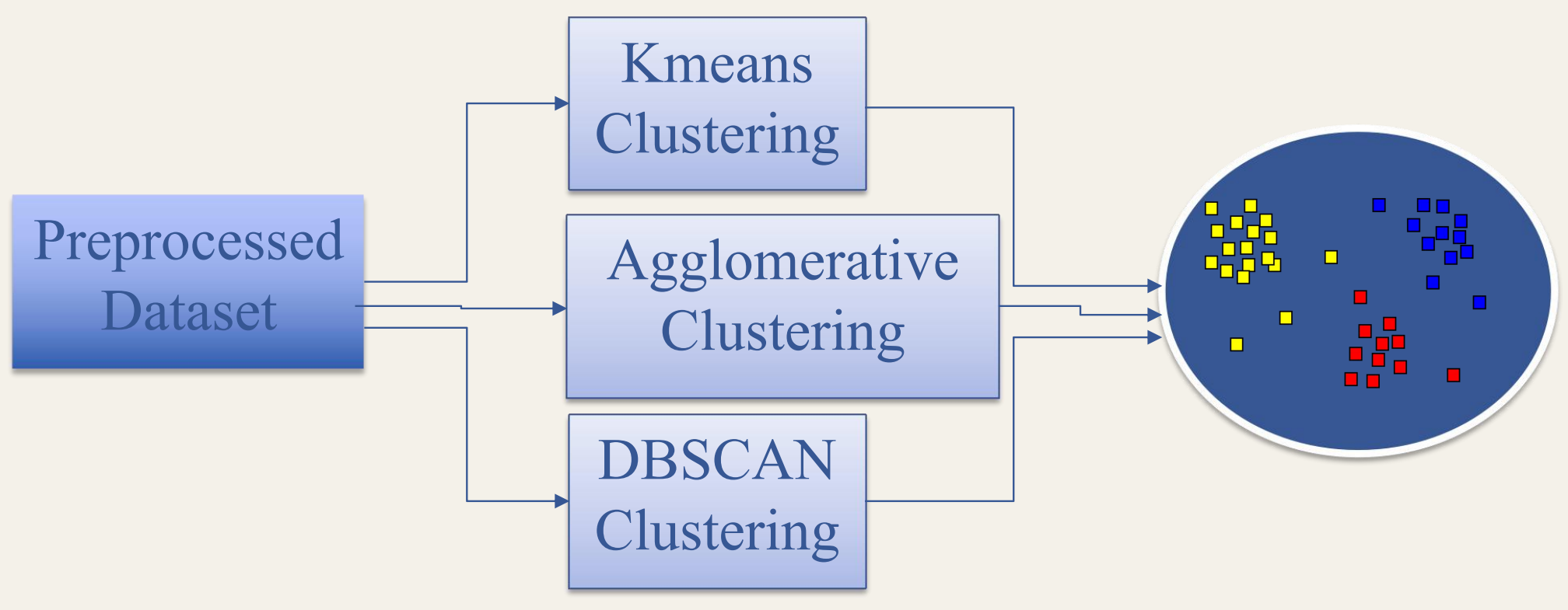
Subset 2: 2001-2021

Dataset Scaling through Standardization

Data Preprocessing

For the K-means algorithm, the number of clusters was set to 3 for both time periods, and the algorithm was fitted to the scaled data using the KMeans() function from sklearn.cluster. Similarly, the hierarchical clustering algorithm was also set to 3 clusters for both time periods using the AgglomerativeClustering() function from sklearn.cluster.

For the DBSCAN algorithm, the DBSCAN() function from sklearn.cluster was used with eps set to 1 and min_samples set to 3 for both time periods.


Preprocessed Dataset

Kmeans Clustering

Agglomerative Clustering

DBSCAN Clustering

Figure 3: Methodology

Results

The evaluation of the clustering results was performed using several methods and The results suggest that there may be a difference in the clustering structure of the data between the time periods of 1981-2000 and 2001-2021. First, the within-cluster sum of squares was computed for K-means clustering.

Table 1: K-means within-cluster sum of squares (WCSS) and Silhouette Score Results

	K-means (1981-2000)	K-means (2001-2021)
WCSS	365956.4273683884	332891.7360758913
Silhouette	0.5696678752499178	0.5815792454499045

The within-cluster sum of squares (WCSS) values for K-means clustering indicate that the clustering for the more recent time period (2001-2021) has a lower WCSS value, which suggests that the clustering is tighter and more compact compared to the earlier time period (1981-2000).

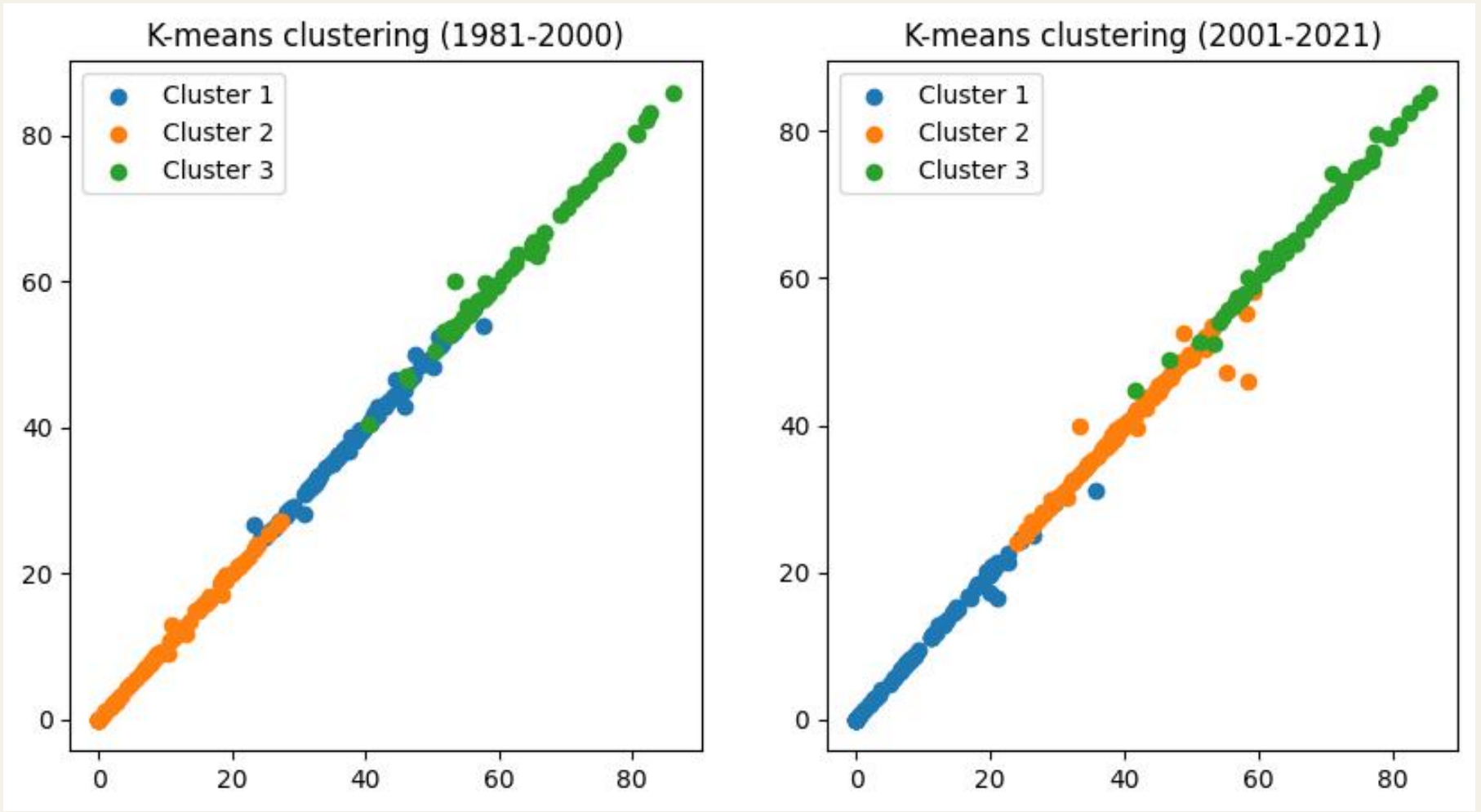


Figure 4: K-means Clustering Results

Second, the silhouette score was computed for all three clustering algorithms.

Table 2: Hierarchical and DBSCAN Silhouette Score Results

	Hierarchical (1981-2000)	Hierarchical (2001-2021)	DBSCAN (1981-2000)	DBSCAN (2001-2021)
Silhouette	0.511253587	0.5317470	-0.44757577	-0.437975692

The silhouette scores for both K-means and hierarchical clustering show that the clustering results for the more recent time period have higher scores than the earlier time period, indicating that the clustering results are better defined and more separated.

However, the silhouette score for DBSCAN is negative for the both time series, which suggests that the clustering structure for this algorithm is not be appropriate for this data. Therefore, it may be better to use K-means or hierarchical clustering.

Overall, these results suggest that there may be a change in the clustering structure of the data over time, and different clustering algorithms may perform differently depending on the time period.

Cluster Visualization

Visual inspection of the clusters was done by creating scatterplots for all three algorithms.

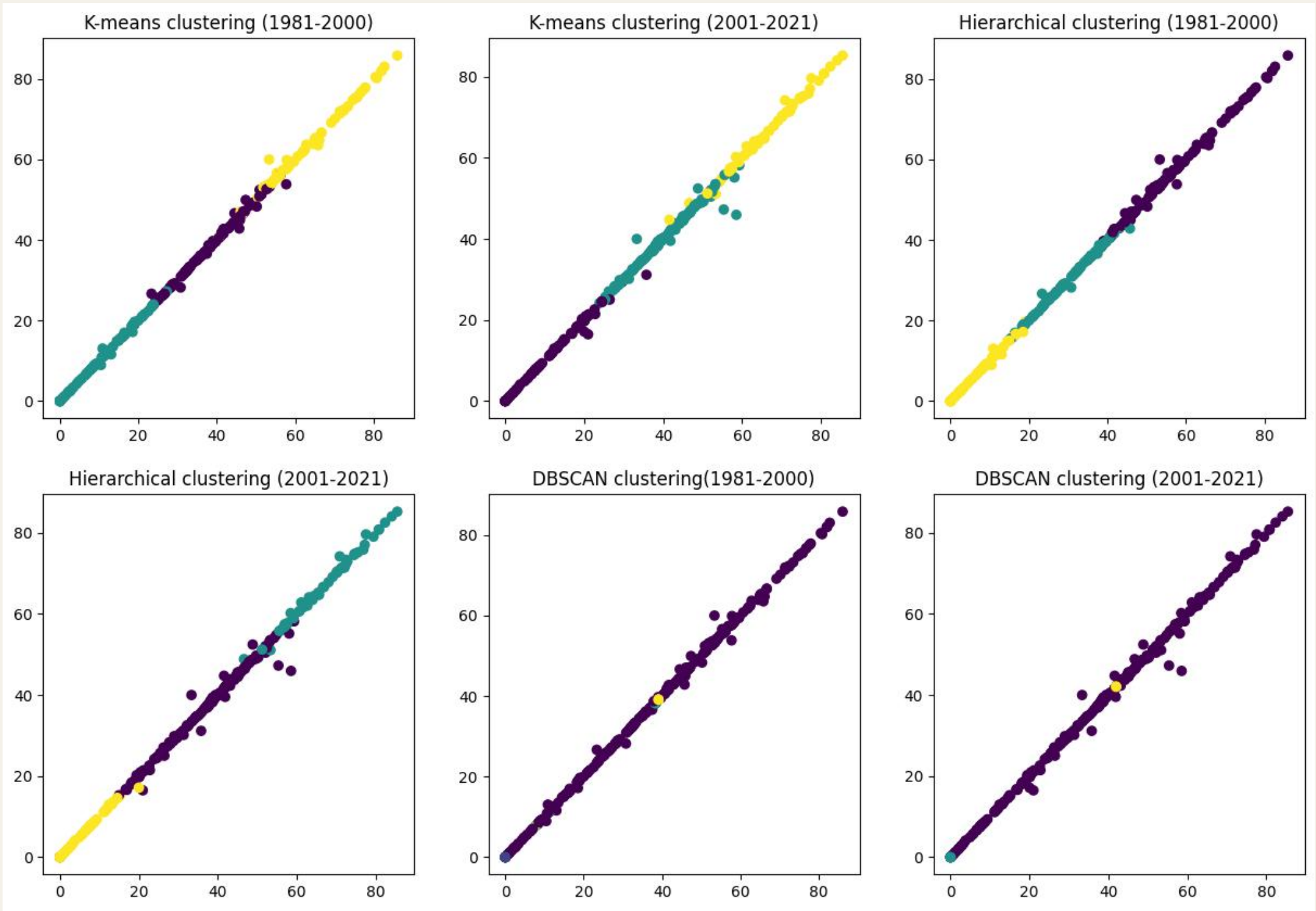


Figure 5: Clustering Results' Visualization

The results indicate that the countries with the highest percentage of land area used for agriculture are primarily located in Sub-Saharan Africa, while the countries with the lowest percentage of land area used for agriculture are primarily located in Europe and North America.

Exponential Growth (Curve) Fitting

Finally, the code performs curve fitting using the exponential growth function. The exponential growth function is commonly used to model the growth of populations, disease outbreaks, and technological innovations. The curve fitting is performed using the curve_fit() function from the SciPy library. The curve_fit() function finds the optimal parameters that minimize the difference between the observed data and the predicted values from the exponential growth function.

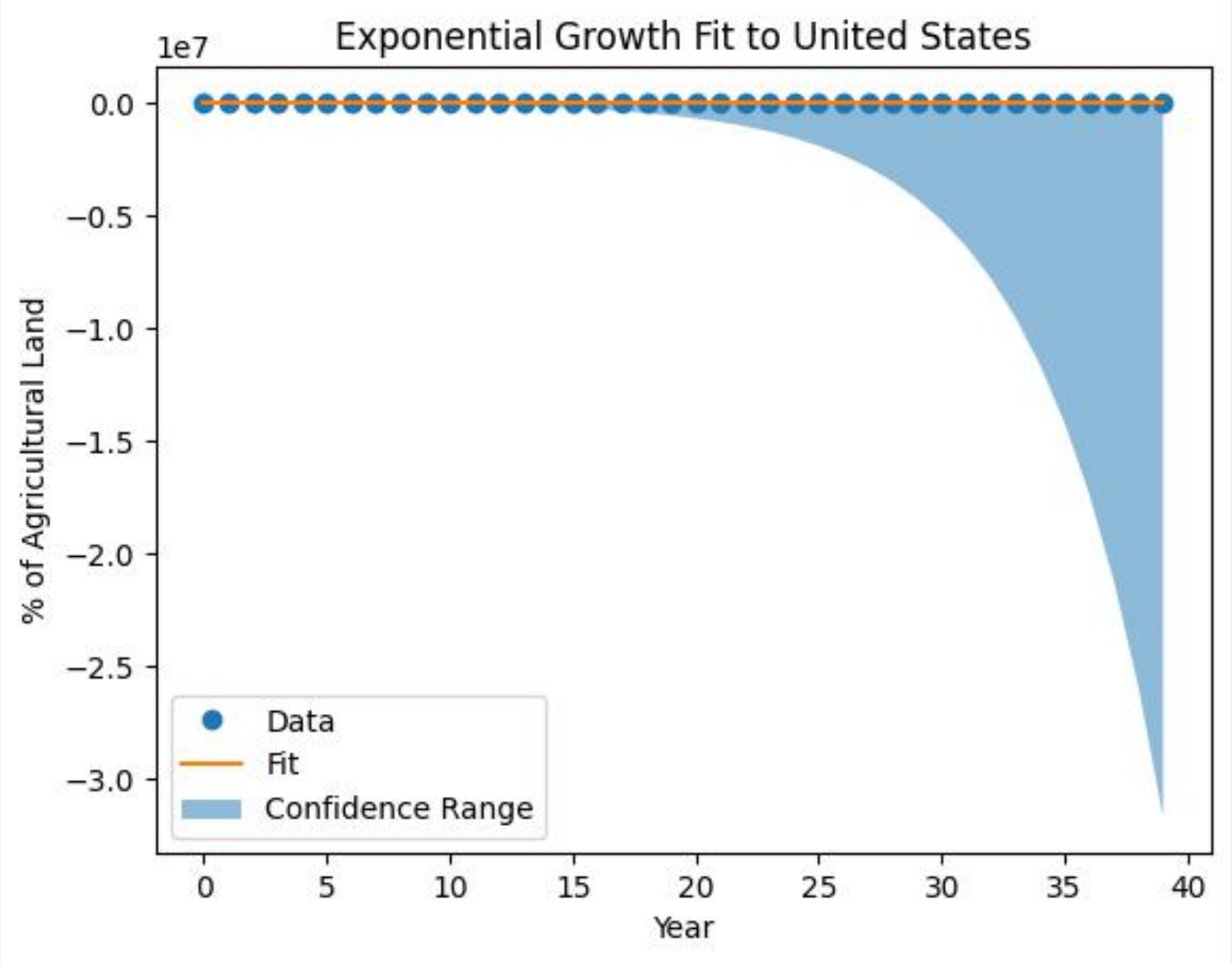


Figure 6: Exponential Growth fitting and prediction

Conclusion

These findings have substantial consequences for policymakers, stakeholders, and researchers. The clustering study can offer policymakers with a better knowledge of the pattern of agricultural land use across different regions and assist in identifying areas where policy actions may be required to promote sustainable agriculture. The analysis can give stakeholders with significant information on the market potential of various regions and assist them make investment decisions in agriculture. For researchers, the analysis can serve as a starting point for future investigation into the underlying causes that drive agricultural land use patterns across various nations and areas.

Nonetheless, this analysis has certain shortcomings. First, the dataset used for this analysis only includes the proportion of land area used for agriculture and does not contain information on the kind of crops grown or the precise agricultural practises employed. Therefore, future research could utilise more detailed datasets to acquire a more comprehensive knowledge of agricultural land use trends. The clustering analysis performed in this work utilised a restricted number of clustering methods and clustering parameters. Additional research could investigate other clustering methods and parameters in order to validate and build upon the findings of this study.

Contact Information

Author's Name	Faisal Zulfiqar
Student ID	22010044
Email ID	faisalzulfiqar1236@gmail.com