Kmeans clustering and fitting for Climate change data from World Bank
Impact of Urban population and other factors

University of Hertfordshire UH
Name: Sanjana Dabbikar
Supervisor: Ralf Napiwotzki

Aim:
The primary aim of this analysis is to employ K-Means clustering to:
Identify natural groupings within climate data.
Understand the characteristics that define each cluster.
Discover relationships and trends that are not immediately apparent.
Provide a clear segmentation that can be used to tailor climate change mitigation and adaptation strategies.

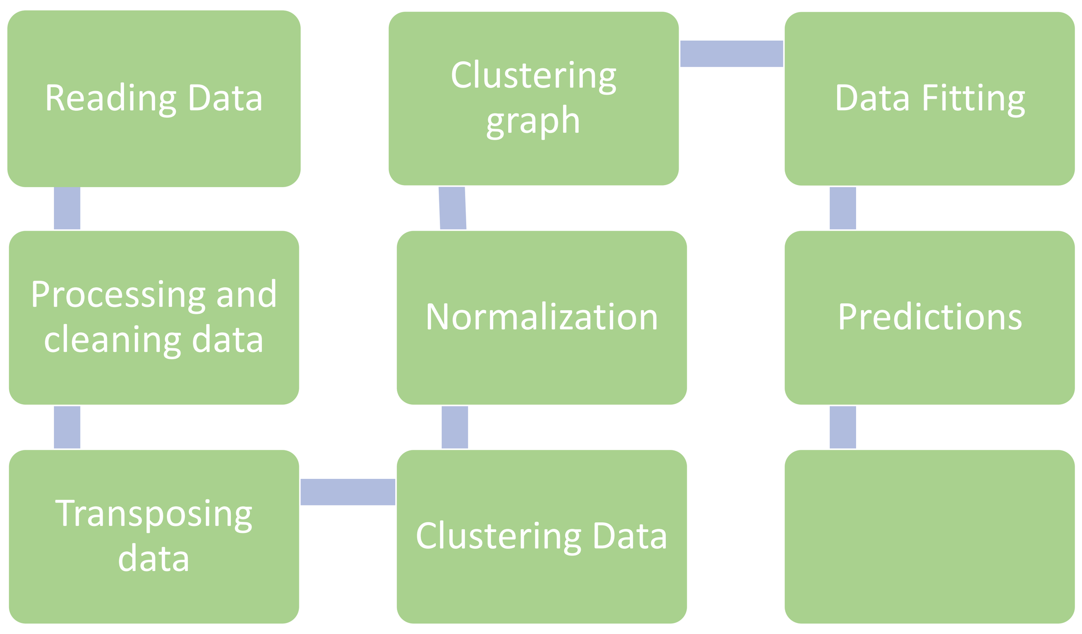

Fig A .Methodology followed

Data Collection: Compilation of a comprehensive dataset from credible sources, including temperature records, emission statistics, and other relevant environmental metrics.
Data Preprocessing: Cleaning, normalizing, and standardizing the data to prepare it for effective clustering.
Feature Selection: Determining the most impactful features that contribute to climate change and are suitable for clustering.
K-Means Clustering: Applying the K-Means algorithm to the processed data, using an iterative approach to partition the dataset into K distinct, non-overlapping subsets or clusters.
Cluster Analysis: Analyzing the properties and centroids of each cluster to interpret the characteristics of the grouped data points.
Validation: Evaluating the quality of the clusters using metrics such as silhouette score, within-cluster sum of squares, and cross-validation with known classifications where available.

Abstract

This report outlines the application of K-Means clustering, a machine learning algorithm, to classify and analyze climate change data. The objective is to identify distinct groupings or patterns within the data that could assist in understanding the multifaceted dimensions of climate change. By segmenting the data into clusters, we aim to reveal insights that could inform policy decisions, scientific research, and public awareness initiatives.

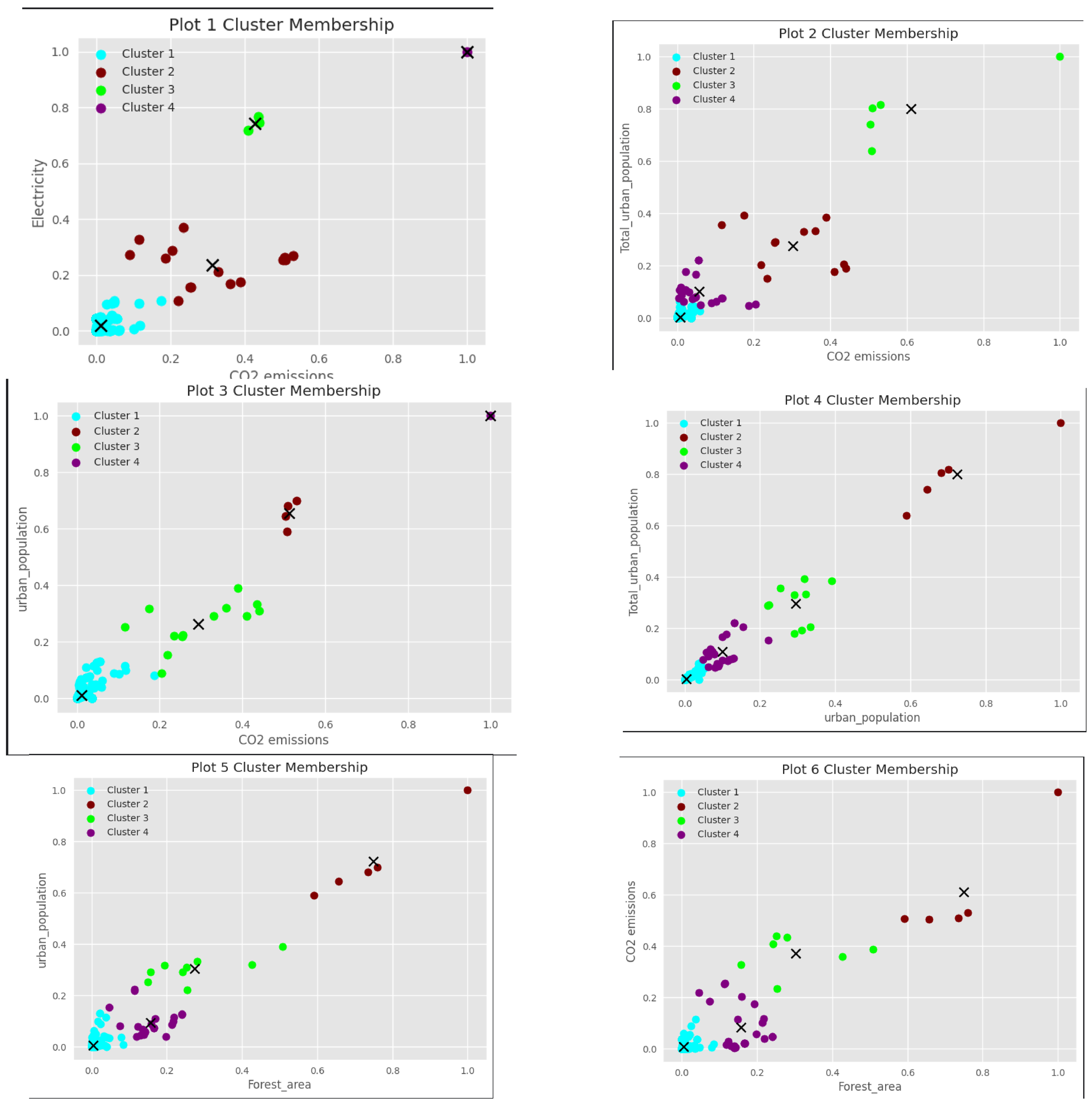

Fig B : Kmeans Clustering

Plot 2: Cluster 1 with low CO2 emissions and total urban population is distinctly separate, suggesting a segment with lower urbanization and industrial activity.
Clusters 2 and 3 show higher CO2 emissions with varied total urban population, possibly indicating more developed or industrial urban areas.
Plot 3: Cluster 1 represents a group with minimal CO2 emissions and low urban population, potentially less developed or rural areas.
Cluster 3 shows a correlation between rising urban population and CO2 emissions, indicating urban development linked with higher emissions.
Plot 4 : Clusters 1 and 3 show low total urban population levels with varying CO2 emissions, which may reflect different stages of urban development and industrialization.
Cluster 2 is sparse, with higher CO2 emissions and total urban population, possibly representing the most urbanized and industrialized areas.
Plot 5 : Cluster 1 shows a group with the lowest urban population and varying forest area, perhaps areas with less human impact.
Clusters 2 and 3, with mid to high urban population levels and lower forest areas, suggest urban expansion impacting forest coverage.
Plot 6 Analysis: Cluster 1 displays the lowest CO2 emissions with a wide range of forest areas, indicating potentially sustainable practices or less industrial activity.
Clusters 2 and 3, with higher CO2 emissions and less forest area, may reflect industrial activities leading to deforestation or reduced forest land.

Results:
The K-Means algorithm identified several clusters within the climate change data. Each cluster was characterized by unique features such as emission levels, geographic regions, economic factors, and temporal changes. The clustering results offered a new perspective on how certain variables are grouped together and how they might influence each other.

China's population is forecasted to potentially stabilize, ranging from 960 million to 1.85 billion by 2050, reflecting a deceleration in growth. In contrast, India's population is expected to continue rising steadily to between 566 million and 802 million by 2050, with less variability in projections. These forecasts highlight the need for strategic planning in both countries to address the demographic challenges and resource management for the coming decades.

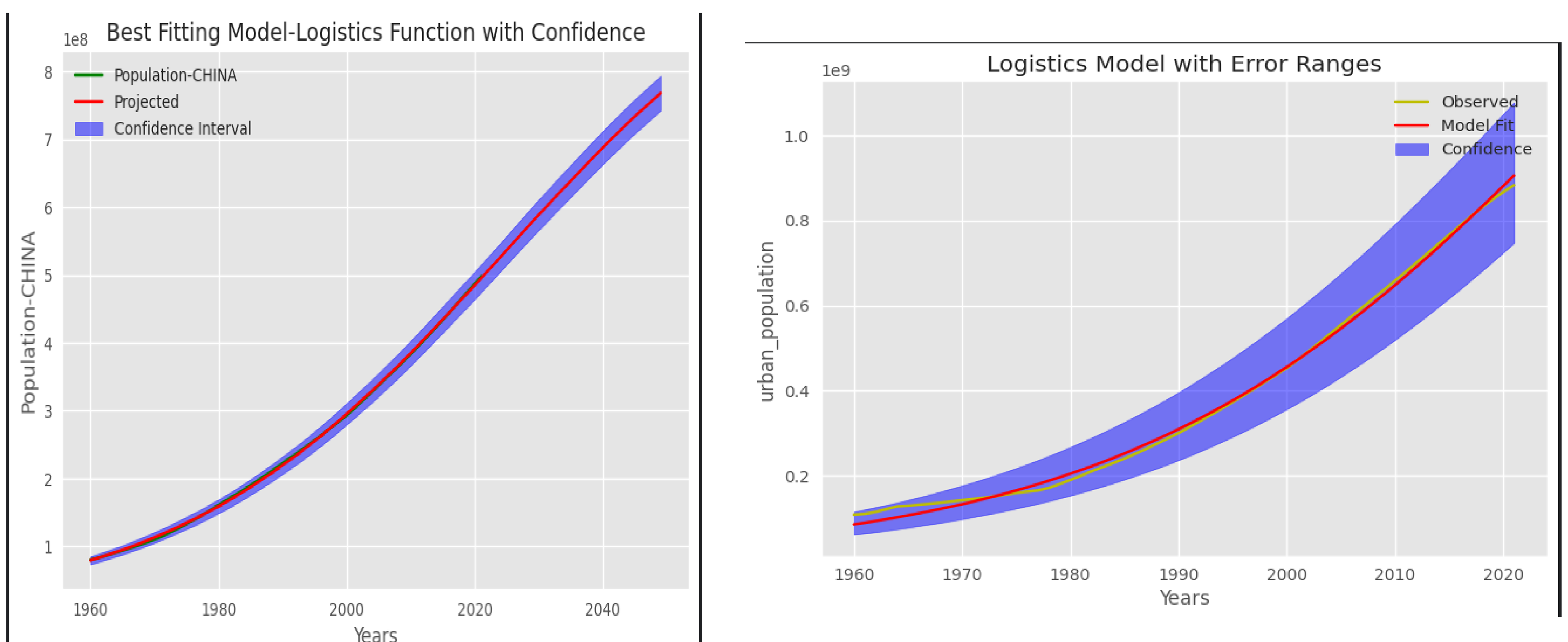

Fig C Fitting of Data with confidence

Forecasted population of India
2030 between 566499545.6749606 and 611564412.5783675
2040 between 663628274.7443433 and 712764221.9663302
2050 between 751018408.6657892 and 802301178.2961738

Forecasted population of China
2030 between 960233818.037207 and 1331234171.243894
2040 between 1193268961.669168 and 1608347539.6547587
2050 between 1406579515.346375 and 1846740475.5274408

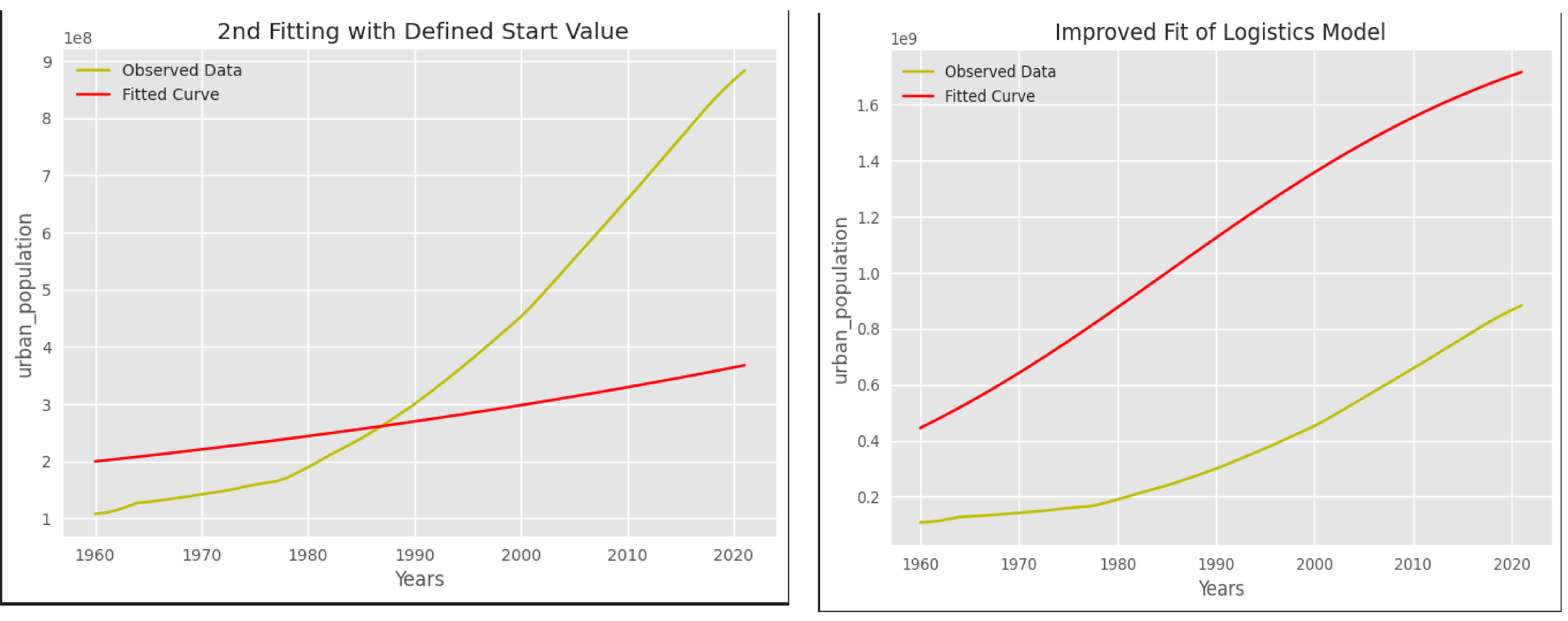

Fig d Fitting of Data with exponential and logistic model

Comparative Insights:

India's population growth rate seems to be steadier and less variable than China's across the decades, possibly due to differences in demographic factors and national policies.
China's wider forecast range indicates greater uncertainty or variability in its future population dynamics.
Both nations will need to plan for significant populations in the coming decades, with implications for resources, infrastructure, and environmental impact.

References
"Population forecasting was performed using logistic growth modeling techniques, similar to methods implemented in statistical software such as Python's SciPy library (Virtanen et al., 2020).“,"The demographic data for modeling purposes were sourced from the National Bureau of Statistics of China, which provides historical population figures (National Bureau of Statistics of China, 2021)."
"The demographic transitions and their implications on population stabilization were reviewed as per the theories discussed in 'The Methods and Materials of Demography' by Siegel and Swanson (2004)."