Project Report

On

AN ANALYSIS OF MARINE FISHERIES PRODUCTION – A COMPARATIVE STUDY IN KERALA AND GUJARAT

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

By

RAMSHEELA P B

(Reg No. SM23AS012)

(2023-2025)

Under the Supervision of

TANIA P R

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI – 682011

APRIL 2025

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **AN ANALYSIS OF MARINE FISHERIES PRODUCTION – A COMPARATIVE STUDY IN KERALA AND GUJARAT** is a bonafide record of the work done by Ms. **Ramsheela P B** under my guidance as partial fulfilment of the award of the degree of **Master of Science in Applied Statistics and Data Analytics** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date:

Place: Ernakulam

Tania P R

Assistant Professor.

Department of Mathematics and Statistics St. Teresa's College (Autonomous)

Ernakulam.

Nisha Oommen

Assistant Professor& HOD

Department of Mathematics and Statistics

St. Teresa's College (Autonomous) Ernakulam.

ST. TERESA'S COLLEGE (AUTONOMOUS) DEPARTMENT OF MATHEMATICS AND STATISTICS ERNAKULAM

External Examiners

1. Sangeetha Chandran 30.04.2025

2. Anju N B (Anjujia) 30.04.25

ST. TERESA'S COLLEGE (AUTONOMOUS) DEPARTMENT OF MATHEMATICS AND STATISTICS ERNAKULAM

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of Mrs. TANIA P R, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam

Date:

RAMSHEELA P B

SM23AS012

ACKNOWLEDGEMENTS

I must mention several individuals who encouraged me to carry out this work. Their continuous invaluable knowledgeable guidance throughout this study helped me to complete the work up to this stage.

I am very grateful to my project guide Mrs. Tania P R for the immense help during the period of work.

In addition, the very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to the faculty, teaching, and non-teaching staff of the department and Colleagues.

I am also very thankful to HOD Mrs. Nisha Oommen for their valuable suggestions, critical examination of work during the progress.

Ernakulam

Date:

RAMSHEELA P B

SM23AS012

ABSTRACT

Marine fisheries production in India is a vital component of the country's economy and food security, contributing significantly to its fisheries sector. India has a vast coastline of about 7500 kilometres, which supports diverse marine ecosystems and a rich variety of fish species. Kerala and Gujarat, two coastal states of India, play crucial roles in the marine fish production landscape. In this project the focus was to forecast the future values of marine fishery landings in Kerala and Gujarat for 7 years that is from 2024 to 2030 . For this traditional method ARIMA and a machine learning model Decision Tree Regression was used. The data was analysed to identify the trends, patterns and non seasonal variations. The better model was found by using Root mean square error and Absolute mean error. The growth rate and correlation between the production of marine fisheries from both the states were compared by using K means clustering and findings were made.

ST.TERESA'S COLLEGE (AUTONOMOUS)
ERNAKULAM

Certificate of Plagiarism Check for Dissertation

Author Name	RAMSHEELA P B
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Smt. Tania P R
Department	P.G. Dept of Mathematics & statistics
Acceptable Maximum Limit	20
Submitted By	library@teresas.ac.in
Paper Title	AN ANALYSIS OF MARINE FISHERIES PRODUCTION – A COMPARATIVE STUDY IN KERALA AND GUJARAT
Similarity	10% AI-5%
Paper ID	3412696
Total Pages	50
Submission Date	2025-03-18 16:05:07

Signature of Student

Signature of Guide

Checked By
College Librarian

ST. TERESA'S COLLEGE LIBRARY ERNAKULAM

* This report has been generated by DrillBit Anti-Plagiarism Software

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Indian marine fish production is a key component of the nation's economy and food security and plays a leading role in its fisheries industry. India has a long coastline measuring about 7500 kilometers, harboring rich marine ecosystems and numerous species of fish. Kerala and Gujarat are two Indian coastal states that are among the leading players in the marine fish production situation.

Southwest coast of India's Kerala state is renowned for its thriving fisheries industry. State marine fish production is high as a result of the fertile waters of the Arabian Sea. The state has 590 kilometres of surf-beaten coast. Continental shelf area along this coast is about 40,000 km2 and overlying water is believed to be one among Indian waters' richest. The estimated future annual sustainable yield of Kerala in the depth range of 0-50 m is 0.599 million tonnes (Scariah et al., 1999).

Gujarat, on the northwestern coast of India, is another crucial component of marine fish production. The state benefits by virtue of the bio-productive waters of the Arabian Sea and the Gulf of Khambhat. Fishing has been chosen as primary livelihood source since time immemorial for the people of Gujarat's coastal belt, stretching along 1,600 km with 14 maritime districts viz. Valsad, Navsari, Surat, Bharuch, Anand, Bhavnagar, Amreli, Gir-Somnath, Porbander, Jamnager and Devbhoomi Dwarka, Morbi and Kutch. The main export products were fish frozen fish, shrimp, squid and cuttlefish. (Sharma et al., 2018).

In Gujarat and Kerala, government interventions, sustainable fishery activities, and aquaculture enterprises play key roles in managing and promoting marine fish production.

Fish farming is an essential element to provide nutritional security, food security and employment in India. Demand for food production has increased mainly with the growth of the population of humans. The yield of fish from capture fisheries is reduced significantly though the demand for consumption of fish is rising significantly because the fish is an excellent protein source with other nutritional attributes and knowledge of the general public about the health impacts of the consumption of fish (Pradeep et al., 2021).

Forecasting plays an important part in daily life. Over the last few decades, some methodologies for time series forecasting have been suggested. Among them, ARIMA models and Decision Tree Regression models are used in this project for Forecasting. The limitations of ARIMA model are that it requires a lengthy time series data to give a good forecast. It is linear model by assumption that data are stationary and have limited ability to cope with non-stationarities and nonlinearities in series data (Anuja et al., 2017). The most popular model for time series forecasting is autoregressive integrated moving average (ARIMA). One of the main drawbacks of this model is the linearity assumption. (Yadav et al.,2020). The clustering methods are utilized to study the rise in production over years for both states.

The results of this project will be immensely beneficial to policymakers, fisheries administrators, and Kerala and Gujarat stakeholders. It can be utilized to develop sustainable fishing practices, implement proper conservation measures, and make the fisheries industry sustainable in the long time.

1.1 About the dataset

The data is collected from the official website of Central Marine Fisheries Research Institute (https://www.cmfri.org.in/annual-data). The dataset consists of the yearly data of marine fishery production in Kerala and Gujarat from 1950 to 2023 (74 years).

1.2 Objectives of the study

1. To model and Forecast Marine fisheries production in Kerala and Gujarat using ARIMA Model.
2. To model and Forecast Marine fisheries production in Kerala and Gujarat using Decision Tree Regression Model.
3. To compare both the models and to find the better model using RMSE and MAE
4. To compare the trends and variations in the landings over time using K means Clustering.

CHAPTER 2

LITERATURE REVIEW

Literature Review offers the synopsis of findings reported in the field of Forecasting Marine Fisheries production. Based on the available studies, the findings help to explicate the various methods that can be utilized in the current study.

Scariah et al., (1999) examined pattern of marine fisheries production in Kerala. Fisheries play a very critical role in Kerala's Economy. Even though it contains just one tenth of the country's coastline, Kerala's landing yield is over 30 % of India's aggregate marine fish production. The sector contributes the major source of employment to almost 147900 active fishers and virtually an equivalent number of others working in the activities of processing and marketing. Earnings from Kerala's sea produce exports have risen significantly in the period of 1970 to 1999. Production of marine fisheries is thus one of the prime concerns of the state's economic planners and the target of numerous development schemes. Results based on the data gathered by the Central Marine Fisheries Research Institute, using its well tested and well accepted stratified multi-stage random sampling design have been taken into consideration for this purpose.

Sathianandan et al.,(2006) examined influence of introducing outboard engine crafts on Karnataka and Kerala marine fish production. Outboard sector contribution is the key part of the overall marine fish production from Kerala and Karnataka states. The effect of the intervention is assessed here employing two popular time series techniques applied in intervention analysis. The first technique employs seasonal ARIMA model and the second employs regression model with ARMA type errors. Quarter wise composite production of sea fish from the two states during the period 1960-2000 were used for the impact study. The research reached a conclusion that for Kerala appropriate model found was seasonal ARIMA type model and for Karnataka appropriate model was regression model with ARMA errors. Based on the recent estimated intervention models, intervention' s effect was estimated at 2.26 lakh tonnes and 88 thousand tonnes per year respectively for Kerala and Karnataka.

Mini et al., (2015) Estimated CPUE series for northeast coast of India fishery and compared among Holt-Winters, ARIMA and NNAR models. Quarterly landings data from January 1985 to December 2014 were used in estimation and forecasting the model . The accuracy of the forecast was determined using Mean Absolute Error, Root Mean Square Error and Mean Absolute Percent Error. From the comparison of models, Holt-Winter's model performance was observed to give better forecast compared to Autoregressive Integrated Moving Average and Neural Network Autoregression model. Holt-Winters model with smoothing factors $\alpha = 0.172$, $\beta = 0$, $\gamma = 0.529$ was observed to be the appropriate model. Seasonality in the series is reflected through gamma value.

Anuja et al., (2017) had carried out a study on Tamil Nadu, Ramanathapuram district being one of the leading maritime districts followed by Nagapattinam and Thoothukudi. The aim of the study was to assess the trends in marine fish production of Tamil Nadu. Annual fish production data from 1988-1989 to 2012-2013 were examined using time series method Autoregressive Integrated Moving Average (ARIMA) model and Regression analysis (curve estimation). They found that best ARIMA model for Tamil Nadu marine fish production was ARIMA (1, 1, 1) which have lowest BIC (Bayesian Information Criterion).

Raman et al., (2017) predicted Odisha marine fish production using seasonal ARIMA model. Quarterly total marine fish landings in Odisha for the period 1985 to 2012 has been considered in this study. The aim was to determine the potential intervention impact estimation and short-run forecasting by ARIMA model estimation in two situations, one with intervention being a part of the model and the second with log transformed data. ARIMA model with log transformed data worked better than the model with intervention component according to Akaike information criterion and Bayesian information criterion of model selection. The model was utilized to predict fish landings for 2013-2015.

Mahalingaraya et al., (2018) constructed a statistical model to predict the overall marine fishery production in India. Autoregressive integrated moving average (ARIMA) model is the most popular model used in forecasting time series. Among the biggest flaw of the model is linearity assumption. To model series that have non-linear patterns, the artificial intelligence methods such as Artificial Neural Network (ANN) model heavily utilize. In this paper an effort has been made to predict the raw productivity of

India with the help of ARIMA and ANN models. Based on empirical finding it is revealed that machine learning methods surpassed the ARIMA model performance.

Boruah et al.,(2020) applied Box-Jenkins model to forecast the Inland and Total Fish production growth rate and trend in India for the period 1978-2018. The secondary data was used in this research, gathered from the Ministry of Agriculture and Farmers Welfare, Govt. of India. In 1978-2018, India's Total fish production rose approximately 2306 thousand tons to 12606 thousand tons and Marine fish production approximately 1490 thousand tons to 3688. ARIMA (0, 2, 1) and ARIMA (0, 2, 1) were the optimal models for inland and India's Total fish production as revealed by the finding. Besides, we used the developed model to forecast the inland and Total fish production in India for the next 20 years up to 2038. It led to the finding that the rate of production of Total fish was found higher than the Production rate of inland fish in India.

Roy and Basu (2024) forecast and simulate inland open water (capture), inland closed water (culture), and marine fish production using yearly time series data reported to the Bangladesh Bureau of Statistics (BBS) from 1984–2019. The study has taken into account the Box–Jenkins method of Autoregressive Integrated Moving Average (ARIMA) approach of forecasting fish production from 2020–2029. The optimal models were ARIMA (1,1,0) for closed water (culture), ARIMA (1,1,0) for inland open water (capture), and ARIMA (0,1,1) for marine fish production based on AIC and BIC model selection criteria. The expected value of capture, culture of inland closed water, and fish production of open water inland exhibited an increasing trend for the time interval between 2020-2029. Policymaking based on such findings of studies creates the sensitization among policy decision makers concerning measures to take advantage of national potential fisheries.

CHAPTER 3

MATERIALS AND METHODS

3.1 TIME SERIES ANALYSIS

Time series is composed of data that are gathered, recorded and monitored at fixed periods of time, time intervals referred here will be months, years or quarter. Time series analysis is employed in multiple disciplines such as research study, economics, business etc. It essentially analyses the points obtained over time and identifies inherent patterns and trends. The time series consists of four major components

1. Secular Trend (Tt)

Trend is used to denote long term changes. It indicates certain and fundamental direction of statistical figures with the flow of time. It is smooth, regular and long term drift. It is applied to general direction of a statistical figure to increase or decrease or stay constant. For instance in a series dealing with population or national income an increasing tendency can be observed whereas in data relating to birth or death or illiteracy a decreasing tendency.

2. Seasonal fluctuations (St)

Seasonal fluctuations are those fluctuations which repeat with some frequency in a certain period of one year or less. Climatic conditions, social traditions, religious activities etc are the reasons for Seasonal fluctuations. For instance the number of road accidents was very low during covid-19 because of lockdown and very high prior to covid-19 scenario.

3. Cyclic fluctuations (Ct)

Cyclic Fluctuations are repetitive movements. Cyclic fluctuations move at intervals (or periods) of over one year. Cycles and Business cycles operate upon Business and Economic series. Cyclic movements proceed through various stages such as prosperity, recession, depression and recovery. Through such various stages, the time series change. Cyclic changes is the term that is used to denote these changing patterns. Prices production, demand etc related series undergo such cyclic changes.

4. Irregular fluctuations (It)

Irregular fluctuations are those produced by uncommon, unforeseen and unintentional happenings. Impact of Earthquake, strike, flood etc gives rise to Irregular fluctuations. These happenings bring about abrupt change of affairs from one state to another. Irregular fluctuations are random in nature. Their occurrence cannot be anticipated as opposed to the other part of time series

3.2 TIME SERIES ECONOMETRICS

3.2.1. Time Series Data

A time series data is said to be a series of values of a variable which changes over time. Observations on a time series can be spaced differently. But the intervals should be of the same range during the observed duration e.g. a day, a week, a month etc. As a rule, the time series is supposed to be stationary in time series-based empirical work.

3.2.2. Stochastic Processes

A process is referred to as stochastic, or random, if the set of a variable is collected over a series of time. A stochastic process either can be stationary or nonstationary.

3.2.3. Autoregressive Model

An autoregressive model is a model in which the dependent variable is regressed upon at least one lagged term of itself. If there is one lagged term of itself in an autoregressive model, then it represents a first-order autoregressive stochastic process, or AR(1). In addition, if the model has p number of lagged terms of the dependent variable, then it represents a pth-order autoregressive process, or AR(p).

3.2.4. Stationary Process

There are various stationarity types. Second order stationary, also referred to as weakly stationary, is deemed to be adequate in most empirical literature. A stochastic process is weakly stationary if the mean and variance are constant and covariance is time invariant, i.e. statistics don't vary with time. A white noise process is a specific type of stationary stochastic process.

A stochastic process is said to be white noise if the mean is zero, the variance is constant, and the observations are serially uncorrelated.

3.2.5. Nonstationary Process

A stochastic process is considered nonstationary if its mean, variance, or covariance changes with time. A typical pattern for financial data is a random walk, a kind of nonstationary stochastic process. A random walk is an AR(1) process that can have drift or not, which indicates the existence of an intercept. Regressing Y_t on Y_{t-1} estimates the following

$$Y_t = \rho_{t-1} + u_t$$

and if ρ equals 1, the model becomes what is known as a random walk.

3.2.6. Integrated Process

Integrated of the first order, represented by the symbol $I(1)$, is a nonstationary stochastic process that requires one difference to become stationary. Similarly, a nonstationary stochastic process that requires two difference operations to become stationary is referred to as second-order integrated, or $I(2)$. Moreover, a nonstationary stochastic process that must be differenced d times is referred to as integrated of order d, represented by the notation $Y\text{t} \sim I(d)$. The integrated of order zero, represented by the notation $Y\text{t} \sim I(0)$, is a stationary time series with no differencing.

3.2.7. Deterministic Trend

It is possible to foresee a deterministic time series with full accuracy. However, because of the probability distribution of future values, the majority of time series are partially stochastic and partially deterministic, making complete prediction impossible.. If a variable is dependent on its past values and a time variable, it is estimated by the following;

$$Y_t = \beta_1 + \beta_2 t + Y_{t-1} + u_t$$

where t is a variable that measures time chronologically and u_t is an error term, assumed to be white noise. The equation is known as a random walk with drift and deterministic trend and is stochastic but also partially deterministic, due to the time trend t.

3.3 ARIMA PROCESS

The acronym for Autoregressive Integrated Moving Average is ARIMA. In order to predict future values, the autoregressive component, often known as the autoregressive (AR), employs the variable's historical values. In order to make the time series stationary, the integrated component (I) subtracts each observation from its preceding observation, a process known as differencing. Stationarity is an assumption in time series that guarantees the statistical characteristics of the series remain constant over time. The error term of the time series is modelled by the Moving Average (MA) component, which determines the discrepancy between the observed and anticipated values using the autoregressive component.

3.3.1 Assumptions in ARIMA

1. Data should be stationary- Stationarity is an assumption in time series which ensures that the statistical properties of the series do not change over time. A series with cyclic behaviour and white noise can also be a stationary time series.
2. Data should univariate - Auto regression (AR component) is the regression by the past values since ARIMA model works well with single variate data.

3.3.2 Steps to be followed

Step1: Exploratory data Analysis

Three steps are involved in exploratory analysis: stationarity check, data visualization, and data description. We should be well informed about our data domain, including its size and variables, while describing it. As the name implies, data visualization involves presenting the data using pie charts, line graphs, bar graphs, histograms, and other visual aids in order to investigate more data characteristics**.**

Detecting Stationarity and seasonality: A time series can be determined to be stationary or to have seasonality components using a variety of techniques. Plotting the time series versus time is a visual method used in graphic analysis. The graph's objective is to determine whether the time series exhibits seasonality and trend or whether it meets stationarity standards. One technique for locating the seasonality component in time series data is seasonal decomposition.

Augmented Dickey-Fuller Test:

Augmented Dickey Fuller test is used to test the stationary of a time series.

H_0: The series is non-stationary v/s H_1: The series is stationary

The test statistic is given by,

$$DE_t = \gamma\hat{}/SE(\gamma)$$

Autocorrelation Function:

The autocorrelation function is the ratio between the covariance at a specific lag, generally expressed as lag k, to the variance. At lag k, ρ_k denotes the ACF and is defined as follows;

$$\rho_k = \gamma_k/\gamma_0$$

where γ_k is the covariance at lag k and γ: is the variance. The ACF can be plotted by using a correlogram. In the correlogram, if all or most of the lags are statistically insignificant, there is no specific pattern, constant variance, and the autocorrelations at various lags hovers around zero, the time series could be regarded as stationary. This means that a time series is most likely stationary if the ACF correlogram resembles a white noise process.

Partial Autocorrelation Function:

The Partial Autocorrelation function (PACF) of a given time series $\{Z_t\}$ is the partial correlation coefficient between $\{Z_t\}$ and $\{Z_{t+h}\}$ obtained by fixing the effects of

$Z_{t+1}, Z_{t+2},..., Z_{t+h-1}$.

Autoregressive Integrated Moving Average Model

A process Zt is said to be Autoregressive Integrated Moving Average (ARIMA(p, d, q))

If $\nabla^d Z_t = (1 - B)^d Z_t$ is ARMA(p, q).

In general, the model can be written as

$$\varphi(B)\,(1 - B)^d\, Z_t = \theta(B) a_t$$

where $\{Z_t\} \sim WN\,(0, \sigma^2)$.

Step 2: Model Selection

We may have multiple models that are suitable for the data while analyzing a time series. The one that best fits the data is the one we select. The Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Schwartz's Bayesian criteria, Parzen's criteria for autoregressive transfer functions, and others are some of the widely used criteria for model selection. The model selection process in this work makes use of the Akaike Information Criteria (AIC). The model with the lowest AIC value is chosen through a selection process.

Akaike Information Criteria

Assume that a statistical model with m parameters is fitted to the data. Akaike proposed the following information criteria to assess the model.

$$AIC = -2ln(L) + 2m$$

where m is the number of model parameters. The model with minimum AIC is preferred.

Box-Jenkins Methodology

The iterative application of model identification, estimation, and diagnostic testing is known as the Box-Jenkins technique. A class of basic ARIMA models is chosen using data plots, autocorrelations, partial autocorrelations, and additional information. In essence, this involves determining suitable values for p, d, and q. The same method of examining the ACF and PACF signatures at the seasonal delays is used to determine the seasonal parameters. As described in Box-Jenkins (1976), the $\varphi(\Phi)$ and $\theta(\Theta)$ of the chosen model are estimated using maximum likelihood approaches, back-casting, etc. . The Akaike Information Criterion (AIC) is used in the model selection process. The model with the lowest AIC value is chosen. By taking into account the autocorrelations of the residuals (the sequence of residuals, or error, values), the fitted model is examined for shortcomings. Until step three fails to improve the model, these stages are applied iteratively.

Ljung-Box Test

The Ljung-Box test is used to test whether the autocorrelations of a time series are different from zero. The test statistic is,

$$\bar{Q}m = n(n+2)\sum_{k=1}^{m}\frac{\hat{\gamma}2}{n-k}$$

The statistic $\bar{Q}m$ has a finite sample distribution that is much closer to that of $\chi^2_{(m-p-q)}$. The procedure is to reject the null hypothesis of uncorrelated residuals, if the computed value of $\bar{Q}m$ is larger than the chi-square table value for a specified significance level.

Step 3: Forecasting

One of the objectives of analysing time series is to forecast its future behaviour. That is, based on the observation up to time t, we should be able to predict the value of the variable at a future time point.

3.4 DECISION TREE REGRESSION

A decision node may include one or more branches. The first node is called the root node. A decision tree can be constructed using both category and numerical data. Decision trees are a technique that uses previously encountered data with a known class to infer the class of previously unseen data. Because decision trees are descriptive and predictive, easy to build and test, highly accurate, and simple to integrate into computer storage devices, they are widely employed in data mining. Datasets containing outliers, different data types, and nonlinear relationships can all be modeled using decision trees. Examples of how these techniques are used include decision theory, classification, prediction, and clustering. If the objective variable is continuous, regression trees are simply called decision trees; if the target variable is categorical, classification trees are called categorical trees. The steps involved in building a decision tree remain the same in either case.

As a statistical procedure, the decision tree technique starts with building a tree structure and then turning the data in the data set into the tree. Nodes, branches, and root nodes make up decision trees. In the process of creating rules, questions are posed and actions are taken in response to the answers. When answers are combined, new rules are created. The decision about the first variable to create the question is followed by the construction of the root node, which is the first node of the tree structure. Each branch determines "if-then" rules, and the subsequent node is sent out in accordance with the outcome. Until a new question comes up, the tree structure is handled in this manner. Eventually is a class's final node is reached.

3.5 TOOLS FOR COMPARISON

3.5.1. K- Means clustering

An iterative process called K-means divides N objects into K distinct clusters. Perhaps the most popular clustering technique is K-means, which is particularly well-known among partitioning-based clustering techniques that use centroids to present clusters. The within-cluster squared error criterion is used to assess the quality of k-means clustering. The K-means algorithm is used to minimize the K-means problem. It has several variations, which will be covered in more detail below. However, in order to apply any of these variations, it is necessary to know how many clusters are present in the data; it will take several runs or trials to determine the ideal number of clusters. Since the inclination to generate a global optimum depends on the size, number of variables, and properties of the data set, there is no best k-means algorithm. The two iteration phases of the k-means clustering methods are the assignment or initialization phase, which uses an iterative process to assign each data point to its closest centroid using the Euclidean metric, and the centroid update phase, which updates the clusters' centroids based on the partition that was determined by the previous phase. When no data point change clusters or a predetermined maximum number of iterations are reached, the iterative process comes to an end. The algorithm is based on the minimization of the average squared Euclidean distance between the data points and the cluster's centre known as centroid, where centroid is the centre of a geometric object and it is seen as a generalization of the mean. The Euclidean distance formula is given by

$$||x - x'|| = \sqrt{\sum (x_d - x'_d)^2}$$

Where x and x' are the data points and x_d and x'_d are the corresponding values .

3.5.2 Root Mean Square Error and Mean Absolute Error

The present study uses Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)for comparison of models. Root Mean Squared Error (MSE) is the square root of average value of squared difference between actual and predicted values. MAE is the average of the absolute differences between predicted and actual values. Both RMSE and MAE can be used to compare the difference model performance on certain data.

$$RMSE = \sqrt{(1/n * (y_i - \hat{y}_i)^)} \qquad MAE = 1/n * \Sigma|y_i - \hat{y}_i|$$

CHAPTER 4

ANALYSIS AND RESULTS

4.1 DATA DESCRIPTION

For the purpose of study the annual data of Kerala and Gujarat was collected from official website of Central Marine Fisheries Research Institute (https://www.cmfri.org.in/annual-data). The data consist of yearly landings of Marine Fishery landings from the year 1950 to 2023.

4.2 INTRODUCTION

This analysis chapter of this study involves two important techniques used for forecasting and understanding patterns in data. We use two methodologies; ARIMA Modelling and Decision tree Regression. ARIMA is the method that takes into account the non seasonal patterns in time series data. In this model it uses the historical trends and patterns and makes future predictions. In ARIMA modelling we assume that the data is stationary and univariate. Another technique used in analysis is Clustering in order to identify the trends and years which show similar production in both years. K-means clustering is used for this.

4.3 ARIMA MODEL

4.3.1 ARIMA MODEL OF KERALA

4.3.1.1 Time series plot of fish landings

The initial step in time series is to draw a time series plot. The time series plot of fish landings of kerala from 1950 to 2023 is given below.

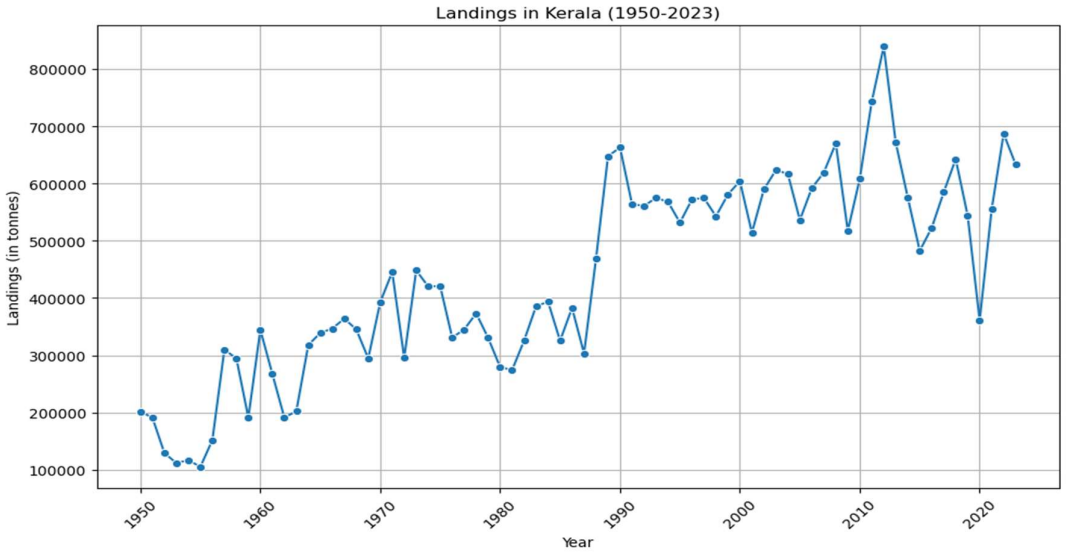

Figure 4.3.1: Time series plot of Landings -Kerala

4.3.1.2 Decomposition of time

The second step is to perform seasonal decomposition to capture the trend, seasonal and random components of time series. Figure given below depicts the seasonal plot.

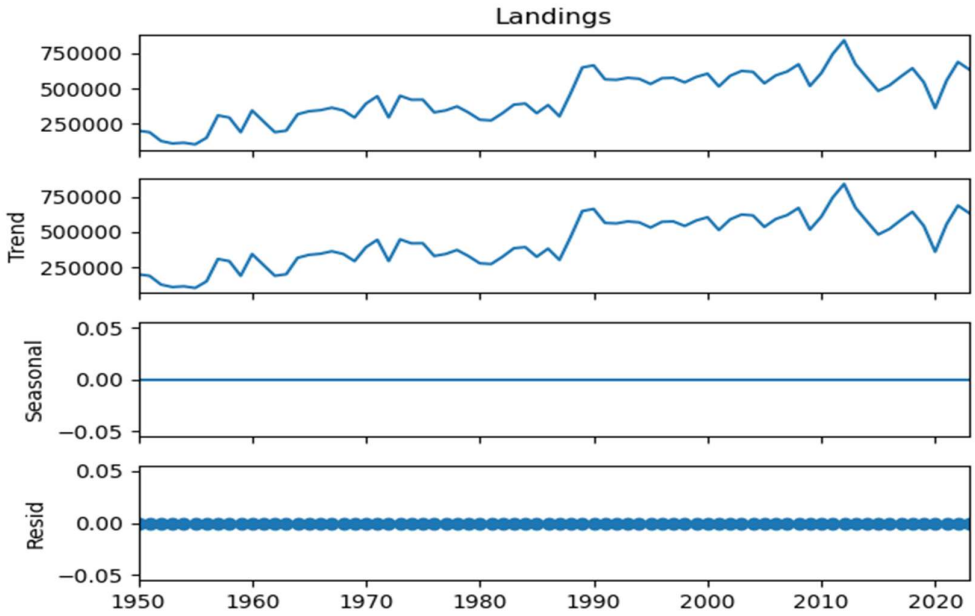

Figure 4.3.2 : Decomposition of time-Kerala

From the figure it is clear that the data has no seasonality.

4.3.1.3 Stationarity check using Augmented Dickey-Fuller Test

To test the time series data for stationarity using ADF test, follows a hypothesis testing approach.

The null hypothesis H_0 is given by,

H_0: The data is non stationary.

The alternative hypothesis H_1 is given by,

H_1: The data is stationary

Results obtained

ADF test statistic: -1.742280, p-value: 0.409471

Critical Values:

1%: -3.5274258688046647

5%: -2.903810816326531

10%: -2.5893204081632653

The p-value is greater than 0.05, so we fail to reject the null hypothesis. The time series is non-stationary.

Hence we perform n order differencing until we get time series stationary in both cases

We perform differencing with n = 1 Now we again check stationarity using ADF test.

Here we test the hypothesis,

H_0: The data is non-stationary.

Against

H_1: The data is stationary

Results obtained

ADF Statistic: -7.666504

p-value: 0.000000

Critical Values:

1%: -3.5274258688046647

5%: -2.903810816326531

10%: -2.5893204081632653

The p-value is less than 0.05, so we reject the null hypothesis. The differenced series is stationary.

Figure given below shows the differenced landings.

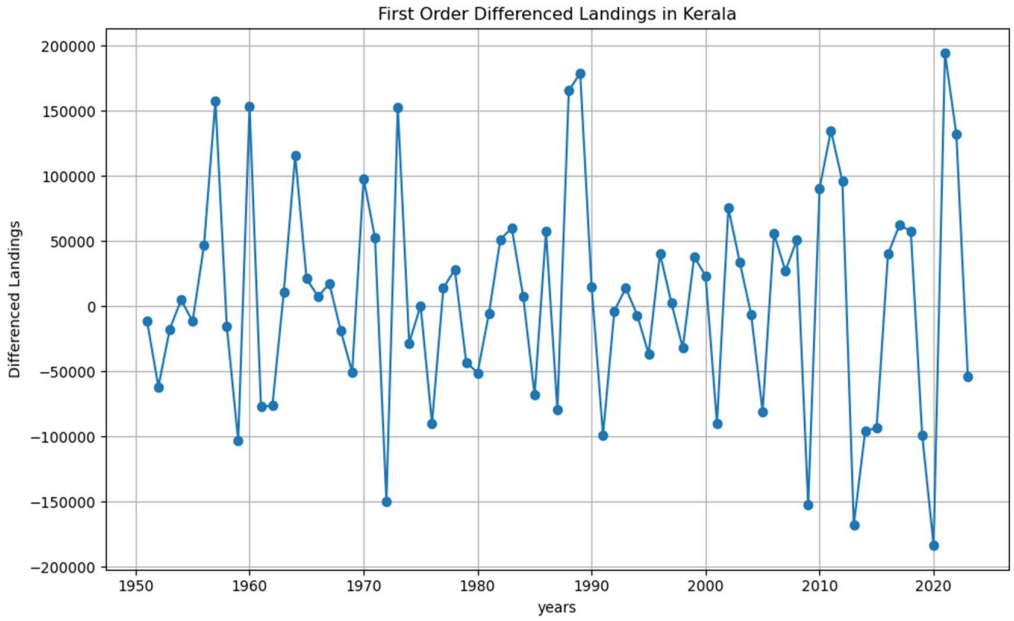

Figure 4.3.3: First order differenced Landings in Kerala

4.3.1.4 Autocorrelation and Partial Autocorrelation Function

Next step in Time Series Analysis is to plot and examine Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). ACF &PACF Plot is given below.

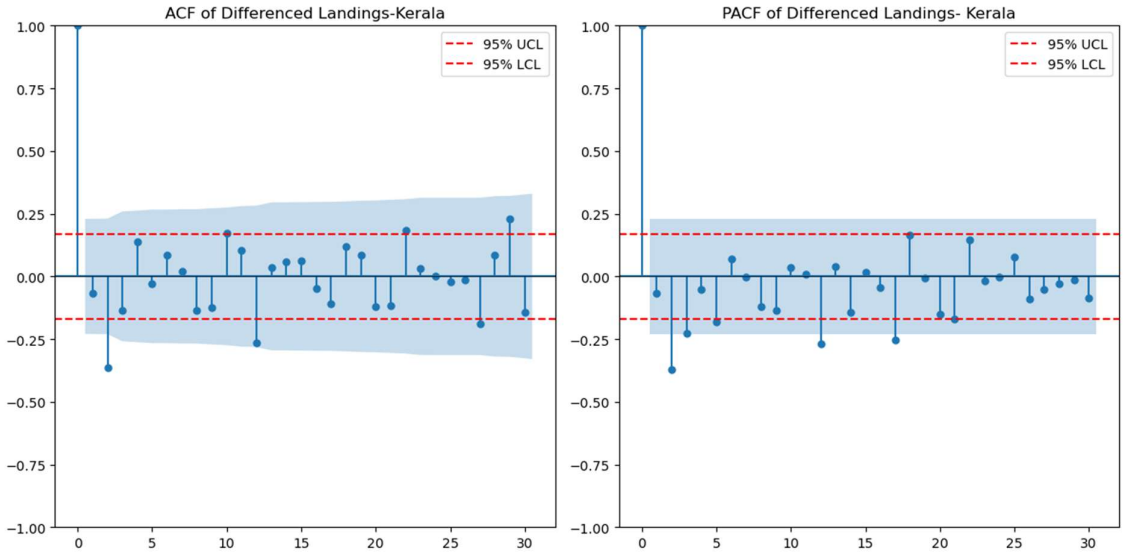

Figure 4.3.4: ACF and PACF of differenced Landings in Kerala

4.3.1.5 ARIMA Model

In this step we choose the best model for forecasting the values. It is done by choosing one model from all possible models according to Akaike Information Criterion (AIC). The model with lowest AIC value is chosen as the best model. Tables below show the possible values of ARIMA model with their AIC value.

Table 4.3.1

SL NO	Model ARIMA (p,d,q)x(P,D,Q)	AIC value
1	ARIMA(0,1,0)x(0,0,0)	1813.716
2	ARIMA(0,1,1)x(0,0,0)	1789.639
3	ARIMA(0,1,2)x(0,0,0)	1756.342
4	ARIMA (0,1,3)x(0,0,0)	1733.027
5	ARIMA(0,1,4)x(0,0,0)	1707.219
6	ARIMA(1,1,0)x(0,0,0)	1815.407
7	ARIMA(1,1,1)x((0,0,0)	1784.096
8	ARIMA(1,1,2)x(0,0,0)	1757.844

9	ARIMA(1,1,3)x(0,0,0)	1729.848
10	ARIMA(1,1,4)x(0,0,0)	1709.206
11	ARIMA(2,1,0)x(0,0,0)	1782.497
12	ARIMA(2,1,1)x(0,0,0)	1780.295
13	ARIMA(2,1,2)x(0,0,0)	1757.643
14	ARIMA(2,1,3)x(0,0,0)	1718.067
15	**ARIMA(2,1,4)x(0,0,0)**	**1693.586**
16	ARIMA(3,1,0)x(0,0,0)	1755.720
17	ARIMA(3,1,1)x(0,0,0)	1757.563
18	ARIMA(3,1,2)x(0,0,0)	1763.632
19	ARIMA(3,1,3)x(0,0,0)	1731.573
20	ARIMA(3,1,4)x(0,0,0)	1696.878
21	ARIMA(4,1,0)x(0,0,0)	1733.168
22	ARIMA(4,1,1)x(0,0,0)	1730.721
23	ARIMA(4,1,2)x(0,0,0)	1983.114
24	ARIMA(4,1,3)x(0,0,0)	1720.442
25	ARIMA(4,1,4)x(0,0,0)	1698.517

Best ARIMA Order: (2, 1, 4)x(0,0,0)

Best AIC: 1693.586

Table 4.3.2

	ar.L1	ar.L2	ma.L1	ma.L2	ma.L3	ma.L4
Coefficients	-0.6702	-0.7963	0.6349	0.5114	-0.6203	0.1949
std error	0.060	0.100	0.198	0.221	0.216	0.212

The ARIMA(2,1,4) Equation is given as

$\Delta y_t = -0.6702\Delta y_{t-1} - 0.7963\Delta y_{t-2} + \epsilon_t + 0.6349\epsilon_{t-1} + 0.5114\epsilon_{t-2} - 0.6203\epsilon_{t-3} + 0.1949\epsilon_{t-4}$.

4.3.1.6 Diagnostic Checking

Diagnostics checking is performed for confirming the validity, effectiveness and reliability of statistical models. The main objective of it is to choose the right and best model.

Ljung-Box test

Null Hypothesis (H_0): The residuals of the model are independently distributed (i.e., no significant autocorrelation remains in the residuals).

Alternate Hypothesis (H1): The residuals of the model are not independently distributed

The Ljung-Box test results of Kerala suggest the following:

Test Statistic : 6.29213

P-value: 0.790152

p=0.790152 is much greater than 0.05, so the residuals appear uncorrelated. It indicates that the fitted ARIMA model adequately explains the autocorrelation structure of the data, and the residuals are likely white noise. Thus, the model is a good fit.

Diagnostic plot is given below

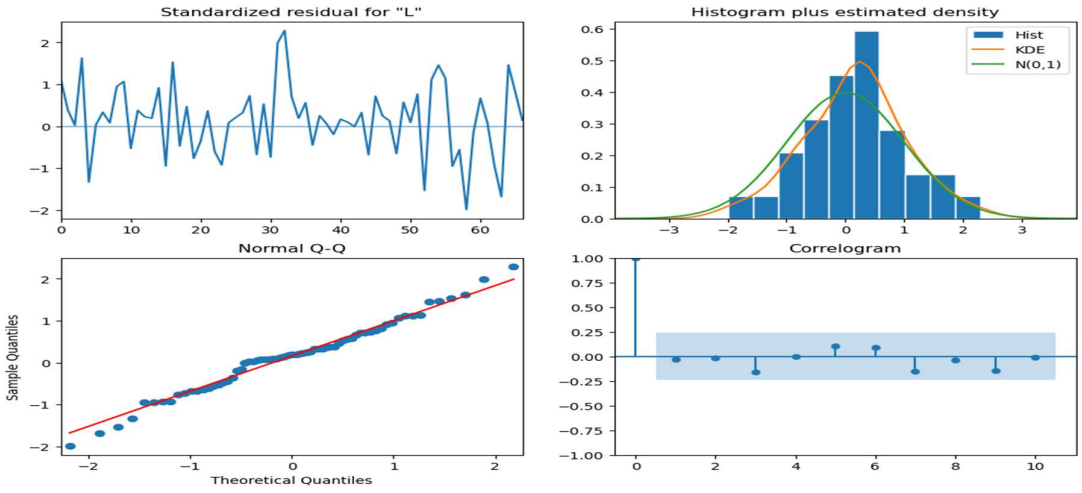

Figure 4.3.5: Diagnostic plot of residuals - Kerala

From Q-Q plot of Kerala, it is clear that most of the residuals are on the same line and standard residual are normally fitted.

4.3.1.7 Forecasting the Sample

Forecasting the Sample means to forecast the actual data points or the training data points. Here we can evaluate model performance on training dataset.

Tables given below is the actual and in sample forecasted values of Kerala

Table 4.3.3

YEAR	ACTUAL LANDINGS (IN TONNES)	PREDICTED LANDINGS (IN TONNES)
2004	616839	606651.736
2005	536215	586413.253
2006	591902	547415.753
2007	619167	612005.036
2008	669982	610479.973
2009	517591	636685.318
2010	608281	522209.702
2011	743123	629410.499
2012	839000	749563.661
2013	671361	745625.244
2014	575644	619243.151
2015	482499	638925.644
2016	522550	534678.136
2017	584686	531923.492
2018	642580	635998.624
2019	543836	617479.944
2020	360867	494125.148
2021	554976	438305.937
2022	686823	622589.707
2023	633258	621538.129

Table given below is the Forecasted future values of Landings

Table 4.3.4

YEAR	FORECASTED LANDINGS (IN TONNES)	LOWER BOUND (95% CI)	UPPER BOUND (95% CI)
2024	556045.698	398740.943	713350.453
2025	597841.893	385405.841	810277.945
2026	613194.932	380762.240	845627.624
2027	567559.893	330118.021	805001.758
2028	585916.985	331507.363	840326.608
2029	609954.746	337252.742	882656.749
2030	579227.382	300038.152	858416.711

Give below is the plot obtained for In-sample and Out-sample Forecasts

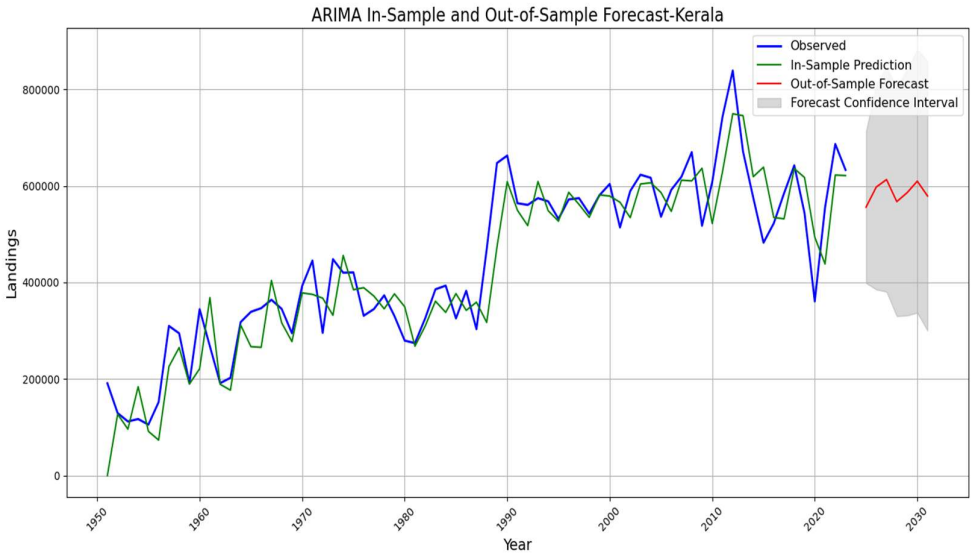

Figure 4.3.6: ARIMA In-sample and Out-sample forecast – Kerala

4.3.2 ARIMA MODEL OF GUJARAT

4.3.2.1 Time series plot of fish landings

The initial step in time series is to draw a time series plot. The time series plot of fish landings of Gujarat from 1950 to 2023 is given below.

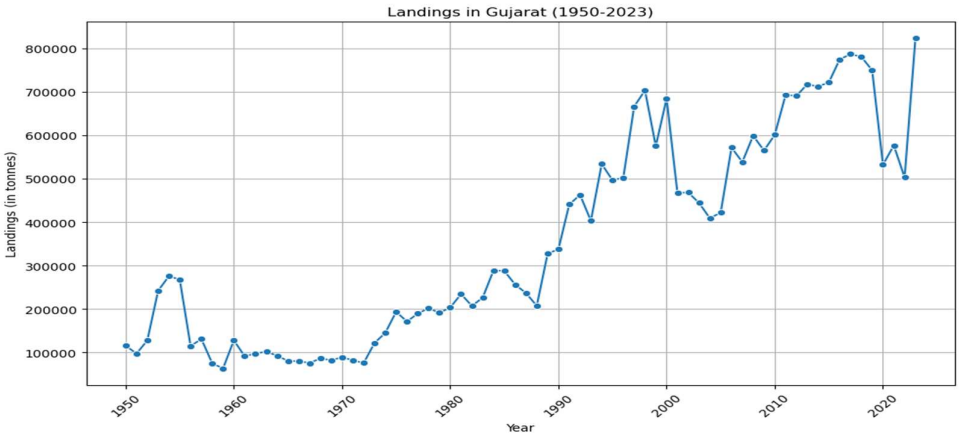

Figure 4.3.7 : Time series plot of Landings- Gujarat

4.3.2.2 Decomposition of time

The second step is to perform seasonal decomposition to capture the trend, seasonal and random components of time series. Figure given below depicts the seasonal plot.

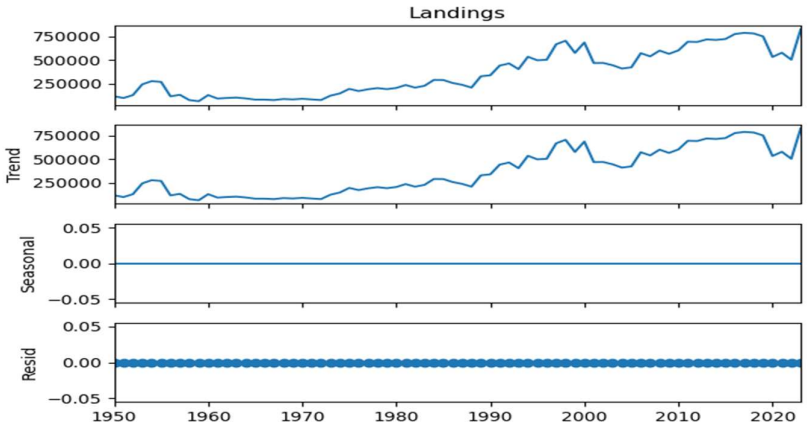

Fig 4.3.8 Decomposition of time-Gujarat

From the figure, it is clear that the data has no seasonality.

4.3.2.3 Stationarity check using Augmented Dickey-Fuller Test

To test the time series data for stationarity using ADF test, follows a hypothesis testing approach.

The null hypothesis H_0 is given by,

H_0: The data is non stationary.

The alternative hypothesis H_1 is given by,

H_1: The data is stationary.

Results obtained

ADF Statistic: -0.183121, p-value: 0.940504

Critical Values:

1%: -3.5319549603840894

5%: -2.905755128523123

10%: -2.5903569458676765

The p-value is greater than 0.05, so we fail to reject the null hypothesis. The time series is non-stationary.

Hence we perform n order differencing until we get time series stationary in both cases

We perform differencing with n = 1 Now we again check stationarity using ADF test.

Here we test the hypothesis,

H_0: The data is non-stationary.

Against

H_1: The data is stationary.

Results obtained

ADF Statistic: -3.476648

p-value: 0.008607

Critical Values:

1%: -3.5319549603840894

5%: -2.905755128523123

10%: -2.5903569458676765

The p-value is less than 0.05, so we reject the null hypothesis. The differenced series is stationary.

Figure given below shows the differenced landings.

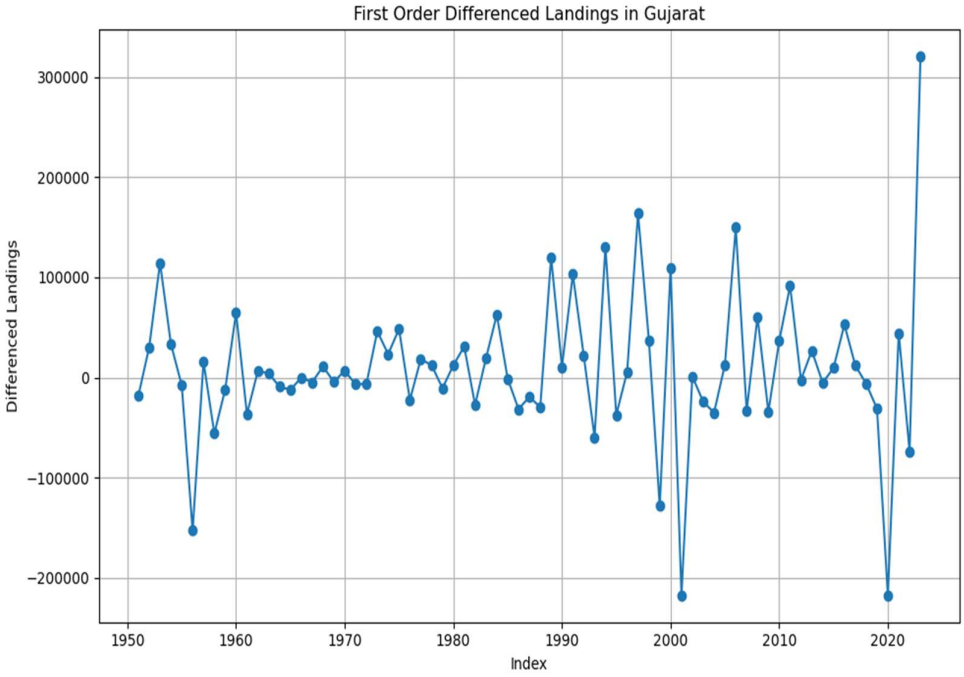

Figure 4.3.9: First order differenced Landings-Gujarat

4.3.2.4 Autocorrelation and Partial Autocorrelation Function

Next step in Time Series Analysis is to plot and examine Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

ACF &PACF Plot is given below.

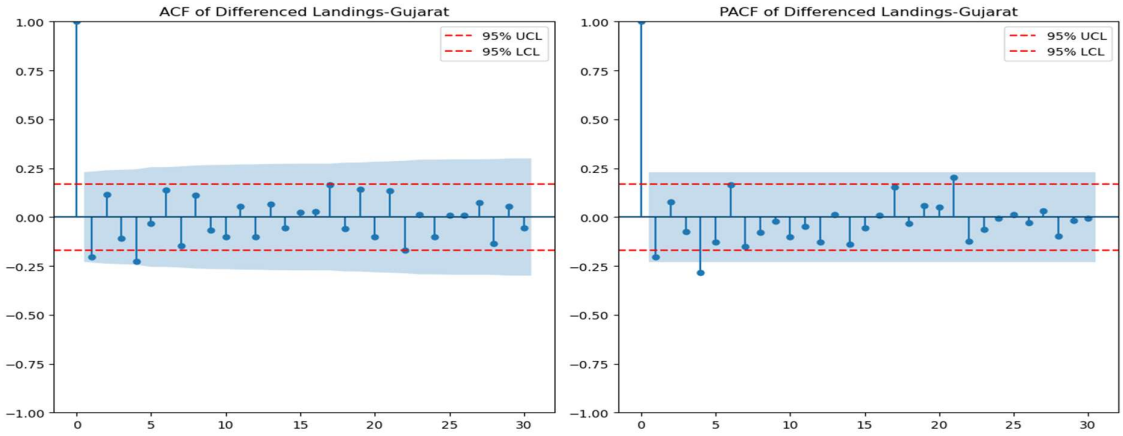

Figure 4.3.10: ACF and PACF plot of differenced Landings - Gujarat

4.3.2.5 ARIMA Model for Landings

In this step we choose the best model for forecasting the values. It is done by choosing one model from all possible models according to Akaike Information Criterion (AIC). The model with lowest AIC value is chosen as the best model. Tables below show the possible values of ARIMA model with their AIC value.

Table 4.3.5

SL NO	Model ARIMA (p,d,q)x(P,D,Q)	AIC Value
1	ARIMA(0,1,0)x(0,0,0)	1800.226
2	ARIMA(0,1,1)x(0,0,0)	1772.580
3	ARIMA(0,1,2)x(0,0,0)	1747.595
4	ARIMA (0,1,3)x(0,0,0)	1718.971
5	ARIMA(0,1,4)x(0,0,0)	1689.914
6	ARIMA(1,1,0)x(0,0,0)	1798.746
7	ARIMA(1,1,1)x((0,0,0)	1772.956
8	ARIMA(1,1,2)x(0,0,0)	1748.268
9	ARIMA(1,1,3)x(0,0,0)	1719.332
10	ARIMA(1,1,4)x(0,0,0)	1690.091
11	ARIMA(2,1,0)x(0,0,0)	1772.764

12	ARIMA(2,1,1)x(0,0,0)	1774.713
13	ARIMA(2,1,2)x(0,0,0)	1741.850
14	ARIMA(2,1,3)x(0,0,0)	1718.824
15	ARIMA(2,1,4)x(0,0,0)	1691.393
16	ARIMA(3,1,0)x(0,0,0)	1749.512
17	ARIMA(3,1,1)x(0,0,0)	1749.171
18	ARIMA(3,1,2)x(0,0,0)	1746.748
19	ARIMA(3,1,3)x(0,0,0)	1719.921
20	ARIMA(3,1,4)x(0,0,0)	1693.180
21	ARIMA(4,1,0)x(0,0,0)	1721.432
22	ARIMA(4,1,1)x(0,0,0)	1722.643
23	ARIMA(4,1,2)x(0,0,0)	1721.369
24	ARIMA(4,1,3)x(0,0,0)	1720.102
25	**ARIMA(4,1,4)x(0,0,0)**	**1689.344**

Best ARIMA Order: (4, 1, 4)

Best AIC: 1689.3447637202764

Table 4.3.6

	ar.L1	ar.L2	ar.L3	ar.L4	ma.L1	ma.L2	ma.L3	ma.L4
Coefficients	-0.6469	-0.0263	0.0004	0.1902	0.5249	0.2391	0.1229	-0.7604
Std error	0.682	0.951	0.623	0.409	0.554	0.715	0.652	0.577

The ARIMA(4,1,4) equation is given as

$\Delta y_t = -0.6469\Delta y_{t-1} - 0.0263\Delta y_{t-2} + 0.0004\Delta y_{t-3} + 0.1902\Delta y_{t-4} + \epsilon t + 0.5249\epsilon_{t-1} + 0.2391\epsilon_{t-2} + 0.1229\epsilon_{t-3} - 0.7604\epsilon_{t-4}$

4.3.2.6 Diagnostic Checking

Diagnostics checking is performed for confirming the validity, effectiveness and reliability of statistical models. The main objective of it is to choose the right and best model.

Ljung-Box test

Null Hypothesis (H_0): The residuals of the model are independently distributed (i.e., no significant autocorrelation remains in the residuals).

Alternate Hypothesis (H1): The residuals of the model are not independently distributed.

The Ljung-Box test results of Gujarat suggest the following:

Test Statistic : 5.219319

P-value: 0.876054

p= 0.876054 is much greater than 0.05, so the residuals appear uncorrelated. It indicates that the fitted ARIMA model adequately explains the autocorrelation structure of the data, and the residuals are likely white noise. Thus, the model is a good fit.

Diagnostic plot of Gujarat is given below:

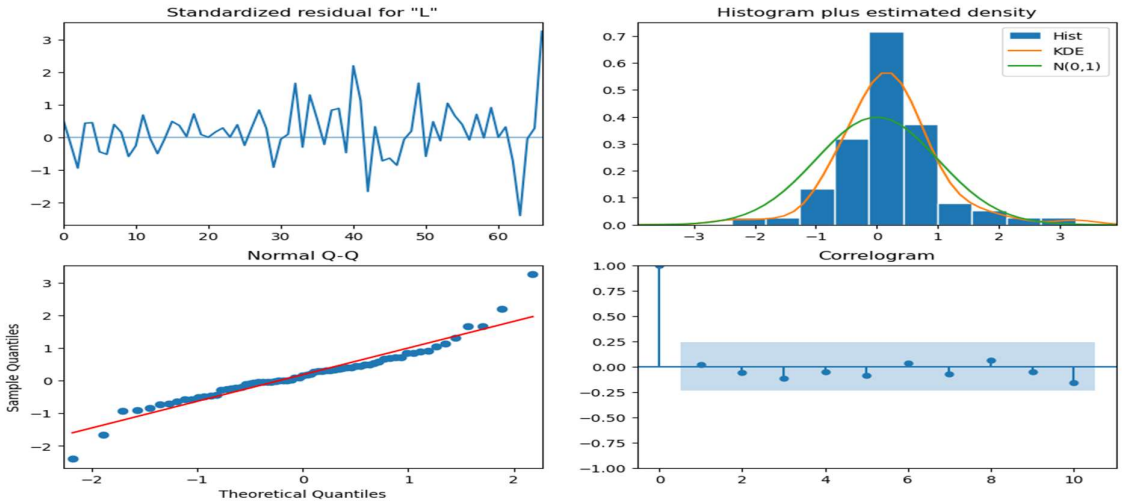

Figure 4.3.11: Diagnostic plot of residuals-Gujarat

From Q-Q plot of Gujarat, it is clear that most of the residuals are on the same line and standard residual are normally fitted.

4.3.2.7 Forecasting the Sample

Forecasting the Sample means to forecast the actual data points or the training data points. Here we can evaluate model performance on training dataset.

Tables given below is the actual and in sample forecasted values of Gujarat

Table 4.3.7

YEAR	ACTUAL LANDINGS (IN TONNES)	PREDICTED LANDINGS (IN TONNES)
2004	408982	413384.601
2005	421873	407546.554
2006	571459	448955.370
2007	538245	580791.880
2008	598813	563119.742
2009	564621	572059.534
2010	601079	523464.954
2011	692702	643565.462
2012	690396	660692.932
2013	717170	722792.112
2014	711930	659355.398
2015	721549	722411.345
2016	774373	706173.164
2017	786495	785890.310
2018	780312	756125.941
2019	749268	617479.941
2020	532031	711368.375
2021	576416	579435.062
2022	502686	481214.135
2023	822786	577891.622

Given below is the Forecasted future values of Gujarat

Table 4.3.8

YEAR	FORECASTED LANDINGS (IN TONNES)	LOWER BOUND (95% CI)	UPPER BOUND (95% CI)
2024	822164.597	674157.336	970171.9
2025	892853.493	702636.550	1083070
2026	834387.206	576194.086	1092580
2027	757103.304	458673.557	1055533
2028	808723.507	497766.602	1119680
2029	791015.789	456648.576	1125383
2030	789807.734	437128.627	1142487

Given below is the plot obtained for In-sample and Out-sample Forecast.

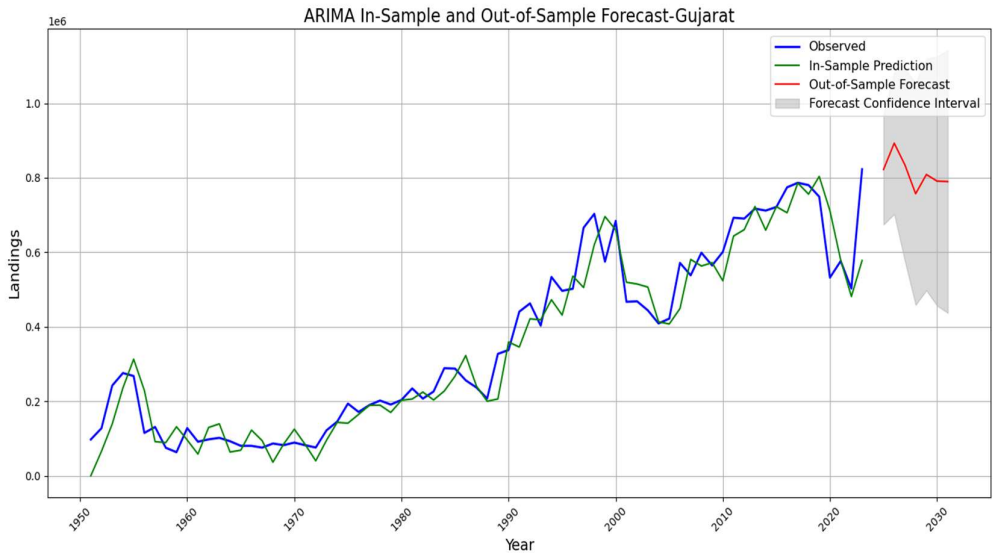

Figure 4.3.12: ARIMA In -sample and Out-sample Forecast of Gujarat

4.4 DECISION TREE REGRESSION

In this analysis Decision Tree Regression was used to forecast the future values of Fish Landings in Kerala and Gujarat. Since the data is univariate we create lag features for Supervised Learning and used the entire data set for training the model.

4.4.1 Decision tree Regression Model of Kerala

In-sample prediction of Landings from 2004 to 2023 is given below.

Table 4.4.1

YEAR	ACTUAL VALUE (IN TONNES)	PREDICTED VALUE (IN TONNES)
2004	616839	590996
2005	536215	550953
2006	591902	590996
2007	619167	601206
2008	669982	590996
2009	517591	550953
2010	608281	590996
2011	743123	743325
2012	839000	839159
2013	671361	681631
2014	575644	590996
2015	482499	513167
2016	522550	580986
2017	584686	601206
2018	642580	654615
2019	543836	513167
2020	360867	360867
2021	554976	554976
2022	686823	654615
2023	633258	648515

Forecasted value of sample is given below.

Table 4.4.2

YEAR	FORECASTED VALUE (IN TONNES)
2024	644261
2025	682984
2026	531776
2027	585624
2028	611563
2029	543991
2030	640610

Given below is the plot of in-sample and out-sample prediction.

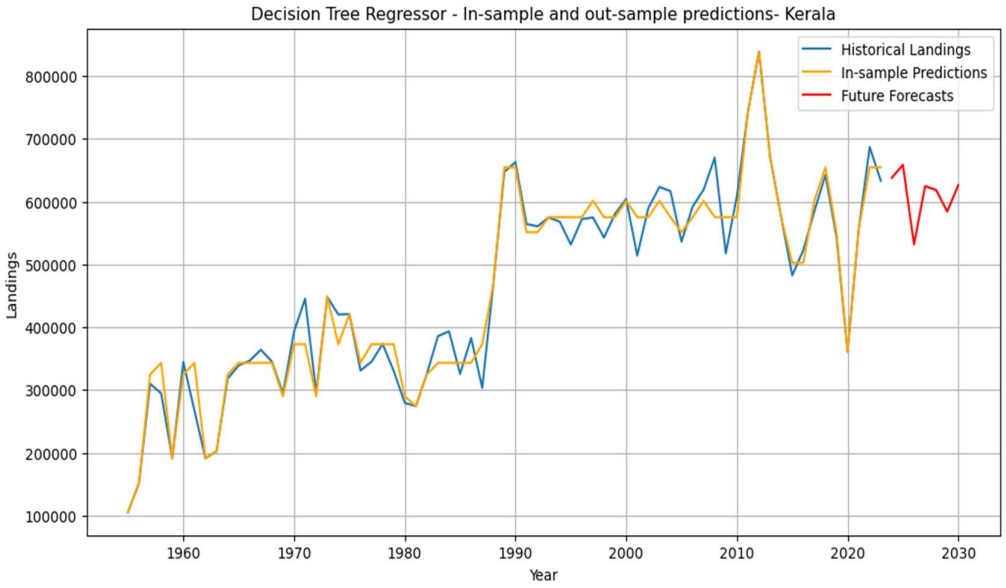

Figure: 4.4.1: Decision tree regressor – In-sample and out-sample predictions- Kerala

4.4.2 Decision tree Regression Model of Gujarat.

In-sample prediction of landings from 2004 to 2023 is given below.

Table 4.4.3

YEAR	ACTUAL LANDINGS (IN TONNES)	PREDICTED LANDINGS (IN TONNES)
2004	408982	408982
2005	421873	421873
2006	571459	571459
2007	538245	517331
2008	598813	645027
2009	564621	648250
2010	601079	637669
2011	692702	655844
2012	690396	757985
2013	717170	768425
2014	711930	755895
2015	721549	745872
2016	774373	755632
2017	786495	757985
2018	780312	745896
2019	749268	754863
2020	532031	532031
2021	576416	576416
2022	502686	502686
2023	822786	757985

Forecasted value of sample is given below.

Table 4.4.4

YEAR	FORECASTED VALUE (IN TONNES)
2024	770529
2025	522324
2026	581354
2027	775738
2028	772022
2029	645686
2030	761016

Given below is the plot of in-sample and out-sample prediction.

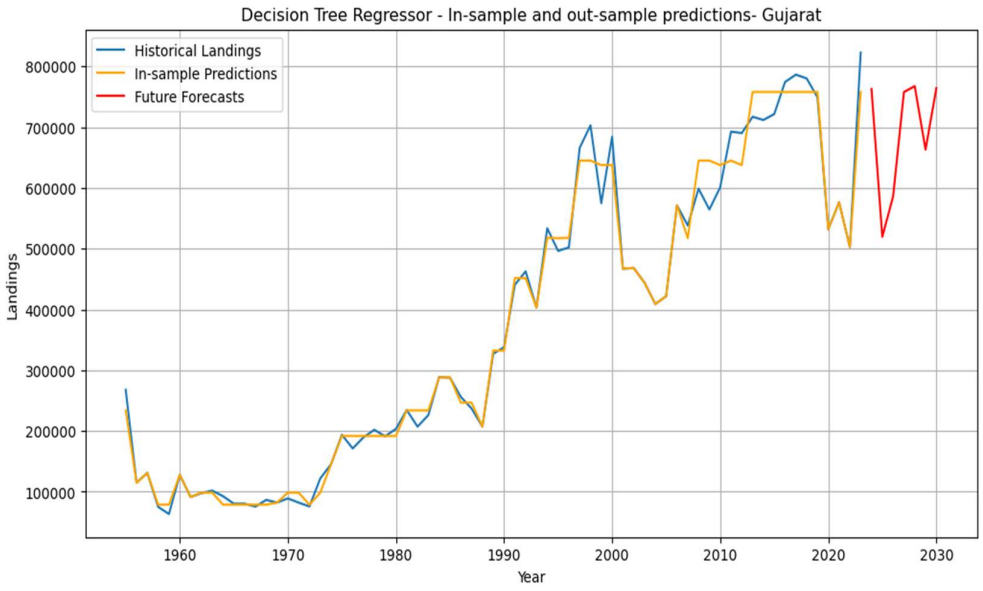

Figure 4.4.2: Decision tree regressor- In-sample and out-sample predictions-Gujarat

4.5 COMPARISON OF RMSE AND MAE VALUES

In order to determine best model from the above two models i.e. ARIMA model and Decisiontree Regression Model, Comparison of the RMSE and MAE values is done for each model. Root Mean Squared Error (RMSE) is the square root of average value of squared difference between actual and predicted values. Mean Absolute Error (MAE) is the average magnitude of the absolute differences between the predicted values and the actual values.

4.5.1 Comparison of RMSE and MAE values of Kerala Models

Table 4.5.1

Model	**RMSE**	**MAE**
ARIMA	68208.577	52254.732
Decision tree Regressor	29077.83	19193.34

4.5.2 Comparison of RMSE and MAE values of Gujarat Models

Table 4.5.2

Model	**RMSE**	**MAE**
ARIMA	63824.832	45752.044
Decision tree Regressor	24609.45	15387.57

From the values it is clear that Decision tree Regressor is the better model for both Kerala and Gujarat with less RMSE and MAE values than ARIMA Models. However when the forecast steps or number of future years increases Decision tree Regressor will give constant values for all years except for the first few. In that case ARIMA model is better for forecasting.

4.6 CLUSTERING AND REGIONAL SEGMENTATION

In this analysis K-means clustering is used to identify years with similar landing patterns in Kerala and Gujarat. By Segment the data into groups, that is high production and low production years and analyse their characteristics can be studied.

4.6.1 Dual axis Time series plot of Landings in Kerala and Gujarat

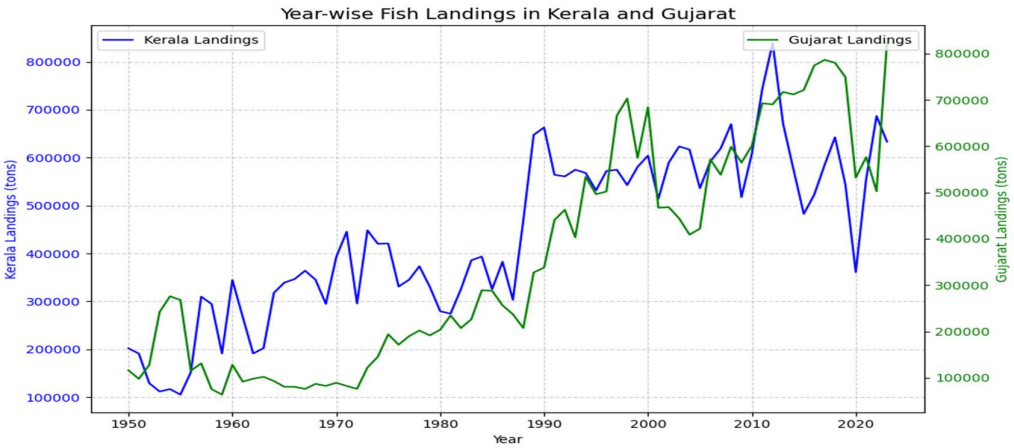

Figure 4.6.1: Year-wise fish landings in Kerala and Gujarat

4.6.2 Determining optimal number of clusters

Using Elbow Method we can determine the value of K in K means clustering or optimal number of clusters. Given below is the plot obtained for Elbow Method.

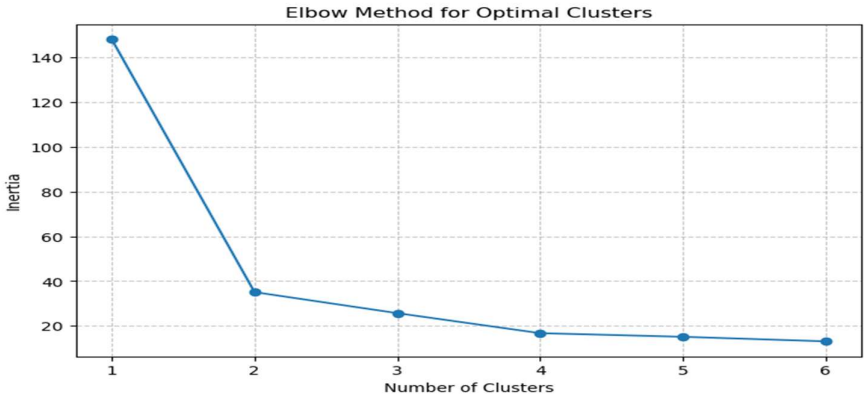

Figure 4.6.2: Elbow method for optimal clusters

From the graph it is clear that inertia is sharply decreases from 1 to 2 and less sharply from 2 to 3 . After 3 clusters decrease in inertia becomes more gradual and diminishes. Hence the optimal number of clusters is likely to be 3.

4.6.3 Clustering of fish landings

Given below is the graph obtained after clustering the datapoints, also years are labelled on it. Correlation between Kerala and Gujarat landings for each clusters are also calculated. Table 4.6.1 is the table containing cluster and their respective years with correlation and Mean landings of both states.

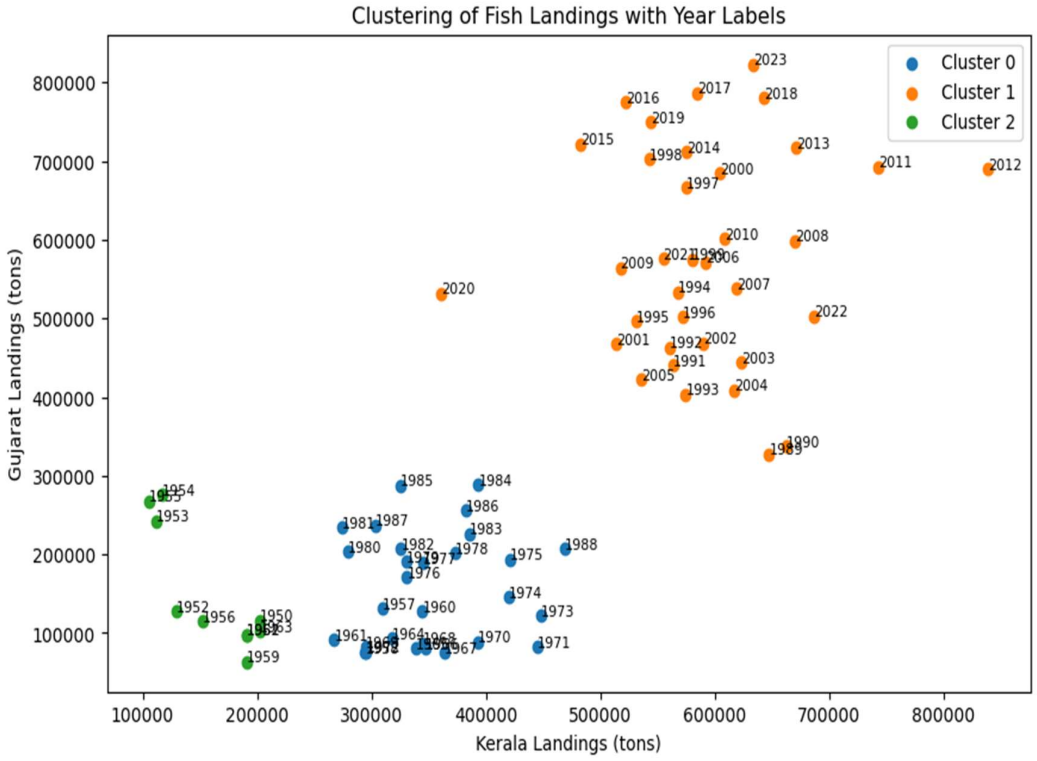

Figure 4.6.3: Clustering of fish landings with year labels

The results are given as a table below.

Table 4.6.1

CLUSTER	YEARS	CORRELATION	KERALA MEAN LANDINGS	GUJARAT MEAN LANDINGS
Cluster 0	1957, 1958, 1960, 1961 1964, 1965, 1966, 1967 1968, 1969, 1970, 1971 1972, 1973, 1974, 1975 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987 1988	0.10	350599	156350
Cluster 1	1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023	0.08	591779	579271
Cluster 2	1950, 1951, 1952, 1953, 1954, 1955, 1956, 1959, 1962, 1963	-0.86	159430	150482

From the above table several inferences can be made related to the production of marine fishery in both states. In cluster 0 there is a weak correlation between Kerala and Gujarat Landings which indicates no linear relationship in their trends during that particular years. Kerala shows significantly higher landings than Gujarat and also Kerala already had an established fishery industry while Gujarat was in development phase during this period.

There is virtually no association between Keralan and Gujarati trends in cluster 1, as indicated by the very modest positive correlation. Gujarat's mean landings (579271 tonnes) are significantly higher than those of clusters 0 and 2. During this time, landings were almost equal, an evidence of Gujarat's fishing industry's rapid expansion. The main causes of Gujarat's fast increase in fish production include regional variations in fish availability, fishing methods, and legislation impacting their fisheries. Even if fig. 4.6.3 shows the impact of COVID-19. Out of cluster 1, 2020 has the lowest production.

In the case of cluster 2 there is a strong negative correlation exist between landings of both the states. Since these are the early years both states had underdeveloped fisheries with low marine total landing. Here when landings in Kerala increased, they decreased in Gujarat, and vice versa. Resource distribution and seasonality differences are the major reason for this.

From the early time (Cluster 2) to the present period (Cluster 1), Gujarat saw rapid growth and overtook Kerala in terms of average landings. Despite slower growth than Gujarat, Kerala was sustaining high and consistent fish landings across all clusters. Low correlations between the clusters suggest that, despite the two states' growth, fishery trends are probably autonomous and are thus influenced by regional elements such ocean biodiversity, policy, and climate.

Approximately 1.74 lakh tonnes were produced in the early 1950s. In the 1980s, it rose to 3.72 lakh tonnes, and in the 1990s, it hit 5.44 lakh tonnes. Improved gear and the introduction of mechanized boats were credited with increasing production. Compared to other marine states, Kerala's fish production increased at a faster rate. Early on, Kerala's fishery development initiatives took off, and they still hold the lead today. The state and federal governments' various agencies provided support for the development initiatives. (K. Balan, 1998)

With a notable 64% increase in marine fish landings over the previous year, Gujarat took first place. With 1.70 lakh tonnes, non-penaeid prawns,mostly little shrimps recorded the state's largest landings in the previous thirty years in 2023. In comparison to 2022, all main resources showed an increase in total landings in Gujarat. With a total of 4.29 lakh tonnes, the multi-day trawlers which have historically been the state's mainstay of marine fishing production recorded the largest landings in the mechanized sector. Landings in the mechanized sector increased by 58% over the prior year. Gir Somnath district, with Veraval fisheries harbour, constituted 51% (4.18 lakh tonnes) of the state’s total landings followed by Junagadh (16%) and Porbandar(14%). (CMFRI-2023)

The fishing industry has been acknowledged as a significant source of revenue and jobs since it fosters the development of numerous ancillary industries and provides affordable, wholesome food. At the same time, it serves as a means of subsistence for a sizable portion of the nation's economically disadvantaged populace. Programs for stock enhancement must be combined with ongoing resource monitoring and fisheries management, which includes protecting habitat, stocking juveniles, and regulating fishing effort appropriately. Therefore, in order to maintain marine resources in the future, marine fisheries require the management and conservation measures required. (Takar and U. R. Gurjar, 2020)

CONCLUSION

In this analysis study of Marine fisheries production in Kerala and Gujarat which are the two main coastal states of India, it was aimed to forecast the future annual production of marine fisheries for the next 10 years that is from 2024 to 2030 and to analyse the trend in the production. For Forecasting two models were used ARIMA(Autoregressive Integrated Moving Average) and Decision tree Regression. From the analysis and by observing the forecasted values obtained from both the models it is clear that the fish production for the next 10 years is increasingly fluctuating for both Kerala and Gujarat.

The third objective of the study was to compare both the above mentioned models by using error metrics RMSE (Root mean square error) and MAE(Mean absolute error). Among the two models Decision tree Regressor has less RMSE and MAE compared to that of ARIMA model. Hence Decision tree Regressor is the better model. However if the forecasting period increases Decision tree regressor give constant values for future production except for the first few years.

The final objective was to identify the trend and variations in the production of both Kerala and Gujarat like a comparison . For that K means clustering is used and classify the years into 3 clusters which have similar production in both the states. From this analysis it is concluded that Gujarat experienced strong growth from the early period to the modern period and caught up to Kerala in average landings. While Kerala was maintaining high and constant fish landings across all clusters, even if growth was not as sharp as seen in Gujarat. Although the two states exhibited growth, low correlations across the clusters indicate that the trends for fisheries are likely independent and therefore driven by local factors such as climate, policy, and biodiversity in the ocean.

REFERENCES

i. Anuja, A., Yadav, V. K., Bharti, V. S., & Kumar, N. R. (2017). Trends in marine fish production in Tamil Nadu using regression and autoregressive integrated moving average (ARIMA) model. *Journal of Applied and Natural Science*, *9*(2), 653-657.

ii. Boruah, B. B., Roy, P., Dutta, A., & Hazarika, B. B. (2020). FORECASTING MARINE AND TOTAL FISH PRODUCTION IN INDIA USING ARIMA MODELS. *Indian Journal of Economics and Business*, *19*(2).

iii. Mahalingaraya, S. R., Sinha, K., Shekhawat, R. S., & Chavan, S. (2018). Statistical modeling and forecasting of total fish production of India: a time series perspective. *Int. J. Curr. Microbiol. App. Sci*, *7*(3), 1698-1707.

iv. Mini, K. G., Kuriakose, S., & Sathianandan, T. V. (2015). Modeling CPUE series for the fishery along northeast coast of India: A comparison between the Holt-Winters, ARIMA and NNAR models. *Journal of the Marine Biological Association of India*, *57*(2), 75-82.

v. Pradeep, M., Ray, S., Tiwari, S., Badr, A. and Balloo, R. 2021. Estimation of Fish Production in India using ARIMA, Holt's Linear, BATS and TBATS Models. Indian Journal of Ecology. 48(5): 1254-1261.

vi. Raman, R. K., Sathianandan, T. V., Sharma, A. P., & Mohanty, B. P. (2017). Modelling and forecasting marine fish production in Odisha using seasonal ARIMA model. *National Academy Science Letters*, *40*, 393-397

vii. Roy, A., & Basu, S. (2024). Forecasting of Fish Production in Bangladeshusing Autoregressive Integrated Moving Average (ARIMA) Models. *BARISHAL UNIVERSITY JOURNAL*, 83.

viii. Sathianandan, T. V., Kuriakose, S., Mini, K. G., & Joji, T. V. (2006). Impact of introduction of crafts with outboard engines on marine fish production in Kerala and Karnataka–a study through Intervention analysis. *Indian Journal of Fisheries*, *53*(3), 271-282.

ix. Scariah, K. S., Devaraj, M., Andrews, J., Seynudeen, M. B., Vijayalakshmi, K., Ammini, P. L., ... & Augustine, S. K. (1999). Production pattern in the marine fisheries of Kerala.

x. Sharma, H. E. M. A. N. T., Swain, M., & Kalamkar, S. S. (2018). Status of Marine Fisheries Sector in Gujarat'. *Agricultural Situation in India*.

xi. Takar, S., & Gurjar, U. R. (2020). Review on present status, issues and management of Indian marine fisheries. *Innovative Farming*, *5*(1), 34-41.