Second Year (Semester-3) Research Assignment on

Time Series Analysis of Marine Fish Sales

in partial fulfilment of the requirement for the successful completion of semester 3 of MSc Big Data Analytics

Submitted By

23-PBD-010

Kotia Divyakumar Ashok

(Semester – III MSc. BDA)

Under the supervision of

Dr. Pravida Raja

2024-2025

Department of Computer Sciences (MSc. BDA)

St. Xavier's College (Autonomous) Ahmedabad – 380009

DECLARATION

I, the undersigned solemnly declare that the research assignment *Time Series Analysis of Marine Fish Sales* is based on my work carried out during the course of our study under the supervision of *Dr. Pravida Raja*. I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that

• The work contained in the report is original and has been done by me under the general supervision of my supervisor.

• The work has not been submitted to any other Institution for any other degree / diploma / certificate in this university or any other University of India or abroad.

• We have followed the guidelines provided by the department in writing the report.

Kotia Divyakumar Ashok

23-PBD-010

MSc. BDA (Big Data Analytics)

St. Xavier's College (Autonomous), Ahmedabad

TABLE OF CONTENTS

1. ABSTRACT

This study extends the work previously done where autocorrelation analysis, time series decomposition, Least Square trend estimation and calculation of seasonal index was carried out for the monthly sales data of marine fish from a fishing vessel in Porbandar district, Gujarat, spanning 2015 to 2022. As a part of this work, two models, ARIMA(1,1,1) and SARIMA(2,1,3)x(0,1,1,12) were fitted to forecast future sales and to better understand the data. The SARIMA model provided a better fit, with lower AIC scores. The findings provide insights for optimizing inventory management and sales strategies in the fishing industry of Porbandar, Gujarat.

Keywords: Marine fish sales, Time Series Analysis, ARIMA, SARIMA,

2. INTRODUCTION

Modeling and forecasting fish catches is crucial for sustainable fisheries management, providing insights that guide decision-making and resource allocation. Research in this area has a long history, with studies like Anuja et al. (2017) forecasting marine fish production in Tamil Nadu using the ARIMA (1,1,1) model. Similarly, Handayani et al. (2020) demonstrated the effectiveness of the SARIMA (21,2,0)(1,0,0,12) model for forecasting catfish sales in Indonesia, while Bako et al. (2013) utilized SARIMA models to predict monthly catches of two fish species over five years.

In this research, we employ the Seasonal Autoregressive Integrated Moving Average (SARIMA) method to analyze marine fish sales data collected from Porbandar district, Gujarat, from 2015 to 2022. The SARIMA model is an extension of the ARIMA framework, specifically designed to incorporate seasonal factors. By using monthly data, this study aims to uncover trends and patterns in fish sales dynamics, which are essential for ensuring the sustainable use of marine resources and supporting the livelihoods of fishing communities.

Understanding these dynamics is vital, especially in a state like Gujarat, where the fisheries sector plays a significant role in the economy. This research contributes valuable forecasts that help policymakers and stakeholders make informed decisions, ensuring both food security and the long-term feasibility of marine resources.

3. REVIEW OF LITERATURE

Anuja et al. (2017) investigates trends in marine fish production in Tamil Nadu, in their study they found that ARIMA (1,1,1) model fits very well for Tamil Nadu marine fish production. Their study highlights the importance of sustainable fishing practices and suggests potential avenues for improving prediction accuracy using advanced methods. Additionally, they acknowledge the limitations of linear models like ARIMA and advocate for the exploration of non-linear methods to capture complex relationships in fisheries data.

Handayani et al. (2020) introduce the SARIMA model for predicting catfish sales, emphasizing its utility in capturing seasonal sales patterns *model named SARIMA (Seasonal Autoregressive Integrated Moving Average) is then proposed to predict the sales. The result shows that SARIMAX (21,2,0)(1,0,0,12) is the best model found in the experiment giving the smallest RMSE.* Their study highlights the practical application of advanced forecasting techniques in the fisheries sector.

Bako et al. (2013) created a model and make predictions using fish catch data. In this study seasonal ARIMA (1, 1,0) $(0,0,1)_{12}$ and SARIMA (0, 1, 1) $(0, 0, 1)_{12}$ models were found fit for monthly catch of two fishes for a period of five years (2007 – 2011) and these models were used to forecast 5 months upcoming catches of *these* fish species. The result will help decision makers to establish priorities in terms of fisheries management.

4. OBJECTIVE, DATA & METHODOLOGY

Objective: The primary aim of this study is to fit the appropriate time series model to forecast the monthly sales of Marine Fish from 2015 to 2022. The objective is to understand how sales change over time, providing valuable information about the marine fish market. By doing so, this research aims to help with fisheries management and policy-making, supporting the development of strategies for using marine resources sustainably.

Data: The data used for this study belong to the sales (in INR) of marine fish caught over a period of 2015 to 2022 by one of the fishing vessels of one of the fishermen of Porbandar district. The data was converted to the .csv format by using the Google lens and was pre-processed using the Python and Excel.

Methodology:

The methodology performed in this study involved the utilization of the Python programming language in which the analysis was mainly performed using the *statsmodels.tsa* library.

a. Checking Stationarity

To analyze the time series data effectively, it is essential to determine if the data is stationary. We utilized the Augmented Dickey-Fuller (ADF) test to check for stationarity. If the data was found to be non-stationary, we applied differencing techniques, specifically first differencing, to stabilize the mean and remove trends from the series. This step was crucial in ensuring the suitability of the data for further analysis.

b. Interpreting the ACF and PACF Plots

After achieving stationarity, we examined the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to identify potential parameters for the ARIMA model. A sharp drop in the ACF after a few lags suggests a potential MA order, whereas a similar pattern in the PACF indicates the AR order. These insights guide the selection of p and q parameters for the ARIMA model.

c. Fitting ARIMA Model

Based on the insights from the ACF and PACF plots, an ARIMA model was initially fitted to the data. The ARIMA model focuses on capturing the relationship between lagged observations (AR), the differencing process (I), and the lagged forecast errors (MA). The ARIMA(1,1,1) model was tested and evaluated for performance using Akaike Information Criterion (AIC)

d. Using Auto ARIMA to select the SARIMA Model with best fit

To account for seasonality in the data, the Seasonal ARIMA (SARIMA) model was explored. Auto ARIMA was used to automate the process of identifying the best combination of model parameters (p, d, q) for both non-seasonal and seasonal components. This led to the selection of SARIMA(2,1,3)x(0,1,1,12) as the best-fitting model based on the lowest AIC value. The model was then used for forecasting future fish sales.

5. DATA AND DATA ANALYSIS

Checking stationarity:

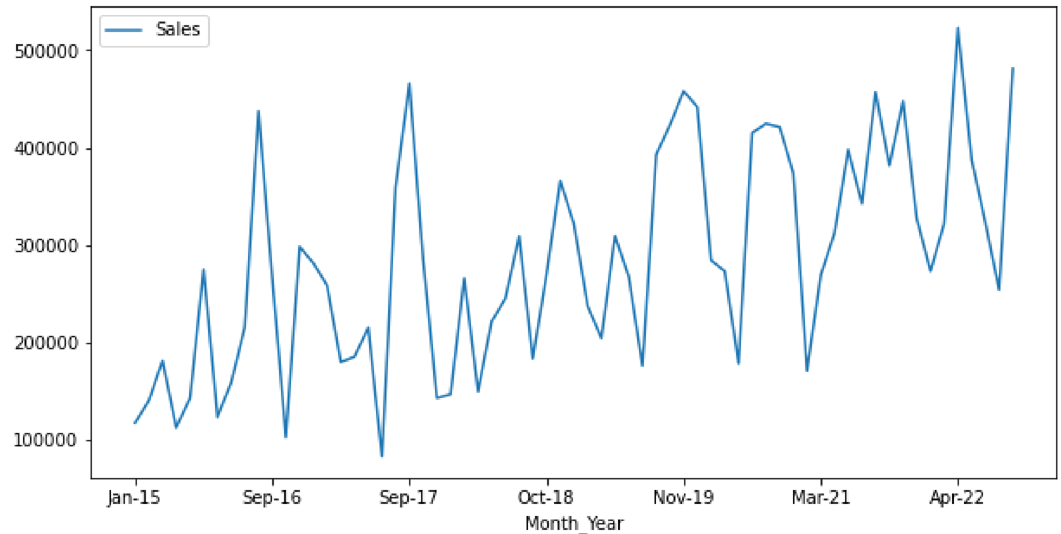

fig 1. Time Series plot of the sales data

From the plot of the data, it clearly seems that statistical properties like variance and mean of the time series are not constant over time. So, we will perform Augmented Dicky Fuller test to be certain that whether or data is stationary or not.

The ADF test works with the following hypotheses:

- Null Hypothesis (H_0): The time series has a unit root (i.e., it is non-stationary).
- Alternative Hypothesis (H_1): The time series is stationary (i.e., it has no unit root).

The results of the ADF test for the original data are as follows:

```
ADF Statistic: -0.7031747202979447
p-value: 0.8459254990846927
The series is non-stationary.
```

fig 2. Result of the ADF test (Original data)

Given that the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating that the series is non-stationary. This suggests the need for differencing to remove the trend and achieve stationarity.

After applying first differencing, the ADF test was conducted again, yielding the following results:

```
ADF Statistic: -5.400921244223636
p-value: 3.37214545697714e-06
The series is stationary.
```

fig 3. Result of the ADF test (First differencing)

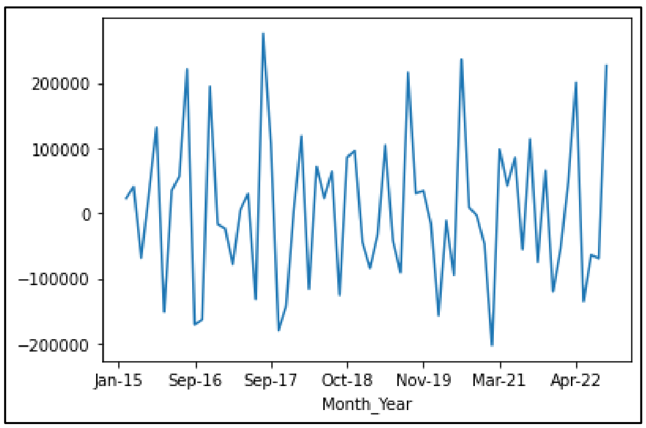

fig 4. Time series plot (After first differencing)

From fig 4., we can say that after the first differencing the trend has been eliminated and also it appears that the seasonality has also been eliminated, still will check it by doing seasonal differencing.

```
ADF Statistic: -2.600707950822828
p-value: 0.09285103666233119
The series is non-stationary.
```

fig 5. ADF test (After Seasonal differencing)

Since the series again became non-stationary this suggest the we have over differenced the series, so we need to consider the series after the first difference only.

ACF and PACF plots

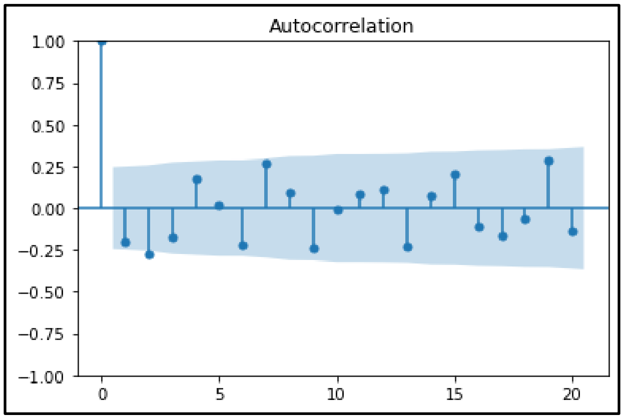

fig 6. ACF plot (after first differencing)

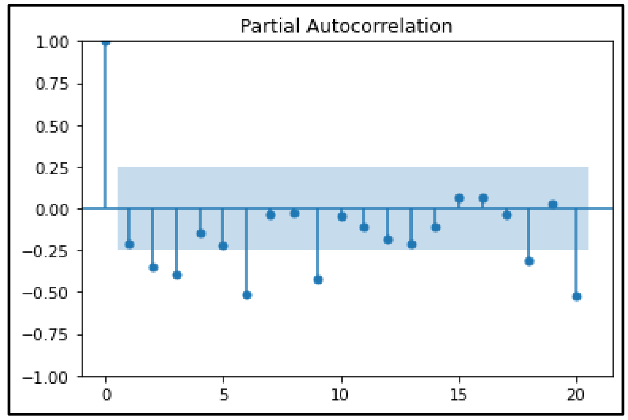

fig 7. PACF plot (after first differencing)

ACF and PACF Plot Analysis

After confirming stationarity through first differencing, the next step was to analyze the ACF and PACF plots.

- **ACF Plot:** The ACF plot displayed a sharp drop-off after the first lag, suggesting the presence of a significant MA(1) component.
- **PACF Plot:** Similar to the ACF, the PACF plot exhibited a sharp cut-off after the first lag, indicating a potential AR(1) component. However, additional spikes were observed at lags 3, 4, 6, 10, 18, and 21. The most notable spikes occurred at lags 6 and 21, although they were smaller than the spike at the first lag. These secondary spikes suggest the possibility of longer-term autocorrelation, but the dominance of the first lag suggests that an AR(1) model might suffice.

Model fitting

Considering the insights obtained from the ACF and PACF plots as discussed above, ARIMA (1,1,1) model was fitted for the data, for which AIC score was obtained as 1657.668.

To account for seasonality in the data, the Seasonal ARIMA (SARIMA) model was fitted whose parameters corresponding to the best fit model were obtained by using the Auto ARIMA method.

```
Performing stepwise search to minimize aic
 ARIMA(0,1,1)(1,1,1)[12]             : AIC=1371.338, Time=0.20 sec
 ARIMA(0,1,0)(0,1,0)[12]             : AIC=1387.600, Time=0.03 sec
 ARIMA(1,1,0)(1,1,0)[12]             : AIC=1382.407, Time=0.10 sec
 ARIMA(0,1,1)(0,1,1)[12]             : AIC=1369.354, Time=0.12 sec
 ARIMA(0,1,1)(0,1,0)[12]             : AIC=1380.860, Time=0.05 sec
 ARIMA(0,1,1)(0,1,2)[12]             : AIC=1371.322, Time=0.31 sec
 ARIMA(0,1,1)(1,1,0)[12]             : AIC=1376.017, Time=0.15 sec
 ARIMA(0,1,1)(1,1,2)[12]             : AIC=1372.386, Time=0.45 sec
 ARIMA(0,1,0)(0,1,1)[12]             : AIC=1375.747, Time=0.08 sec
 ARIMA(1,1,1)(0,1,1)[12]             : AIC=1364.758, Time=0.38 sec
 ARIMA(1,1,1)(0,1,0)[12]             : AIC=1375.060, Time=0.15 sec
 ARIMA(1,1,1)(1,1,1)[12]             : AIC=1366.747, Time=0.48 sec
 ARIMA(1,1,1)(0,1,2)[12]             : AIC=1366.736, Time=0.79 sec
 ARIMA(1,1,1)(1,1,0)[12]             : AIC=1370.884, Time=0.35 sec
 ARIMA(1,1,1)(1,1,2)[12]             : AIC=1367.714, Time=0.97 sec
 ARIMA(1,1,0)(0,1,1)[12]             : AIC=1374.018, Time=0.13 sec
 ARIMA(2,1,1)(0,1,1)[12]             : AIC=1361.255, Time=0.41 sec
 ARIMA(2,1,1)(0,1,0)[12]             : AIC=1372.977, Time=0.22 sec
 ARIMA(2,1,1)(1,1,1)[12]             : AIC=1362.927, Time=0.59 sec
 ARIMA(2,1,1)(0,1,2)[12]             : AIC=1362.755, Time=0.83 sec
```

```
 ARIMA(2,1,1)(1,1,2)[12]             : AIC=1364.898, Time=2.06 sec
 ARIMA(2,1,0)(0,1,1)[12]             : AIC=1368.656, Time=0.15 sec
 ARIMA(3,1,1)(0,1,1)[12]             : AIC=1363.042, Time=0.35 sec
 ARIMA(2,1,2)(0,1,1)[12]             : AIC=1361.903, Time=0.42 sec
 ARIMA(1,1,2)(0,1,1)[12]             : AIC=1362.174, Time=0.22 sec
 ARIMA(3,1,0)(0,1,1)[12]             : AIC=1364.473, Time=0.20 sec
 ARIMA(3,1,2)(0,1,1)[12]             : AIC=1360.570, Time=0.44 sec
 ARIMA(3,1,2)(0,1,0)[12]             : AIC=1375.706, Time=0.21 sec
 ARIMA(3,1,2)(1,1,1)[12]             : AIC=1362.361, Time=0.52 sec
 ARIMA(3,1,2)(0,1,2)[12]             : AIC=1362.226, Time=0.61 sec
 ARIMA(3,1,2)(1,1,0)[12]             : AIC=1367.304, Time=0.36 sec
 ARIMA(3,1,2)(1,1,2)[12]             : AIC=1363.976, Time=1.03 sec
 ARIMA(4,1,2)(0,1,1)[12]             : AIC=1361.331, Time=0.54 sec
 ARIMA(3,1,3)(0,1,1)[12]             : AIC=1354.105, Time=0.66 sec
 ARIMA(3,1,3)(0,1,0)[12]             : AIC=inf, Time=0.56 sec
 ARIMA(3,1,3)(1,1,1)[12]             : AIC=1355.906, Time=0.90 sec
 ARIMA(3,1,3)(0,1,2)[12]             : AIC=1355.789, Time=1.25 sec
 ARIMA(3,1,3)(1,1,0)[12]             : AIC=1360.666, Time=0.88 sec
 ARIMA(3,1,3)(1,1,2)[12]             : AIC=1357.549, Time=2.61 sec
 ARIMA(2,1,3)(0,1,1)[12]             : AIC=1352.467, Time=0.49 sec
 ARIMA(2,1,3)(0,1,0)[12]             : AIC=inf, Time=0.60 sec
 ARIMA(2,1,3)(1,1,1)[12]             : AIC=1354.267, Time=0.75 sec
 ARIMA(2,1,3)(0,1,2)[12]             : AIC=1354.150, Time=1.29 sec
 ARIMA(2,1,3)(1,1,0)[12]             : AIC=1358.935, Time=0.72 sec
 ARIMA(2,1,3)(1,1,2)[12]             : AIC=1355.919, Time=2.38 sec
 ARIMA(1,1,3)(0,1,1)[12]             : AIC=1358.154, Time=0.26 sec
 ARIMA(2,1,4)(0,1,1)[12]             : AIC=1356.161, Time=0.55 sec
 ARIMA(1,1,4)(0,1,1)[12]             : AIC=1357.184, Time=0.61 sec
 ARIMA(3,1,4)(0,1,1)[12]             : AIC=1357.289, Time=1.24 sec
 ARIMA(2,1,3)(0,1,1)[12] intercept   : AIC=1355.435, Time=0.64 sec
```

```
Best Model: SARIMAX(2, 1, 3)x(0, 1, 1, 12) - AIC: 1352.467386305846
```

fig 8. Auto Arima results

From the above mentioned models with different parameters, SARIMA(2,1,3)x(0,1,1,12) was found to be the model that fitted the best with AIC score of 1352.46.

6. FINDINGS AND CONCLUSIONS

In this study, we applied both ARIMA and SARIMA models to forecast fish sales data. The ARIMA model was specified as ARIMA (1,1,1), while the SARIMA model was fitted with parameters SARIMA (2,1,3)(0,1,1,12). The AIC scores for these models were as follows:

ARIMA (1,1,1): AIC = 1366.747

SARIMA (2,1,3) (0,1,1,12): AIC =1352.467

The lower AIC score for the SARIMA model indicates a better fit compared to the ARIMA model. This suggests that the SARIMA model captures the seasonal patterns and trends in the fish sales data more effectively.

The findings align with the study by Anuja et al. (2017), which also found the ARIMA(1,1,1) model to be a suitable fit for fish sales data. This consistency reinforces the reliability of the ARIMA modeling approach in this domain.

From our analysis, it is evident that understanding the sales patterns in the fishing industry is crucial for optimizing inventory and managing supply chains. By employing such time series models, stakeholders in the fishing industry can make informed decisions regarding production and distribution, ultimately leading to improved profitability and sustainability. These insights emphasize the importance of data-driven decision-making in enhancing the operational efficiency of fishing businesses.

7. REFERENCES

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

2. M. Maslim[2], N. Handayani[1], P.Mudjihartono[3] . (2020). Forecasting of Catfish Sales by Time Series. *84 Jurnal Buana Informatika, Volume 11*, 84-93.

3. Hadiza Yakubu Bako[1], Hazel Monica Matias-Peralta[4], Ibrahim Lawal Kane[3], Mohd Saifullah Rusiman[2], Predictive Modeling of Pelagic Fish Catch using Seasonal ARIMA Models. *Agriculture, Forestry and Fisheries.* Vol. 2, No. 3, 2013, pp. 136-140. doi: 10.11648/j.aff.20130203.13