Project Report

On

MODELLING AND PREDICTING OF UNEMPLOYMENT IN KERALA

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

In

APPLIED STATISTICS AND DATA ANALYTICS

By

ANEENA THOMAS

(Register No. SM22AS003)

(2022-2024)

Under the Supervision of

PARVATHY T S

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM,

KOCHI - 682011 MAY 2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

THE FEAR OF THE LORD IS THE BEGINNING OF WISDOM

ST. TERESA'S COLLEGE (AUTONOMOUS) ESTD 1925

CERTIFICATE

This is to certify that the dissertation entitled, STATISTICAL MODELLING AND PREDICTING OF UNEMPLOYMENT IN KERALA is a Bonafide record of the work done by ANEENA THOMAS 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: 29/04/2024

Place: Ernakulam

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

Ms. Parvathy T S

Assistant Professor,

Department of Mathematics and Statistics,

St. Teresa's College (Autonomous), Ernakulam.

Ms. Nisha Oommen

Associate Professor & HOD,

Department of Mathematics and Statistics,

St. Teresa's College (Autonomous),

Ernakulam.

External Examiners

1: .... CHINU JOSEPH.

2: .... LAKSHMI SURESH.

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **Ms. Parvathy T S**, Assistant Professor, Department of 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: 29/04/2024

ANEENA THOMAS

SM22ASOO3

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

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and appreciation to the following individuals who have contributed to the completion of this research work.

Firstly, I would like to thank my supervisor, Ms. Parvathy T S, for her valuable guidance, support, and encouragement throughout the research process. Her insightful comments and suggestions have been instrumental in shaping this research work.

I want to express my heartfelt gratitude to Ms. Nisha Oommen, HOD, Department of Mathematics and Statistics who guided me throughout this journey.

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

Finally, I want to express my gratitude to my family and friends for their constant support and encouragement throughout this journey. Their love and encouragement have kept me motivated and focused on completing this research work. Thank you all for your invaluable contributions to this research work.

Ernakulam

Date: 29/04/2024

ANEENA THOMAS

SM22AS003

ABSTRACT

Unemployment is a critical issue that persists all over the world. It highly affects the economic growth and overall well-being. This project aims to forecast the future unemployment rates of Kerala. This can assist the policy makers to take necessary steps in order to overcome the issue of unemployment.

The data chosen is secondary and consists of two variables namely Date and Estimated Unemployment Rate (%). In accordance with the seasonal decomposition, SARIMA and Holt-Winters models are chosen for forecasting. The two time series models are compared using RMSE (Root Mean Squared Error). The result of the comparison shows that Holt winters is the best model for forecasting the unemployment rates in Kerala.

ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM

Certificate of Plagiarism Check for Dissertation

Author Name	ANEENA THOMAS
Course of Study	M.Sc. Applied Statistics & Data Analytics
Name of Guide	Ms. PARVATHY T.S.
Department	Post Graduate Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	MODELLING AND PREDICTING OF UNEMPLOYMENT IN KERALA
Similarity	5% AI 7%
Paper ID	1663979
Submission Date	2024-04-19 11:35:03

Signature of Student

Signature of Guide

Checked By
College Librarian

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Unemployment arises as an important economic and social challenge affecting not only Kerala, but the whole world. It affects individuals, families and the whole communities. It signifies the unavailability of jobs to be provided to the willing workers so that they could get gainful employment despite their readiness to work. The after effects of unemployment extend far beyond the individual level which highly affects the economic growth and overall well-being.

During these times unemployment persists in multifaceted forms, each one having their own challenges. Rising of shifts in industries and technological progresses lead to structural unemployment leaving the workers unskilled for the available jobs. Cyclical unemployment follows economic cycles which occur due to a negative economic growth. It rises during periods of recession and falls during the times of economic expansion. Frictional unemployment occurs as the workers move between jobs, looking for better opportunities or relocating for employment. Seasonal unemployment occurs due to fluctuations in demand for particular industries based on the time of year.

Households and individuals are also affected in various ways, which becomes a main reason for financial crisis, decreased self-esteem and increased vulnerability to poverty. It also becomes a reason for social exclusion. A very high unemployment rates can obstruct economic growth.

A different approach is required for unemployment that holds various policy interventions, including education investments and skills training, providing support for small and medium-sized enterprises and measures to stimulate growth of economy and creation of job. Furthermore, for sustainable economic growth, developing an inclusive labour market which ensures equal access to opportunities for all segments of society is necessary.

This project titled ‘Modelling and Predicting of Unemployment in Kerala’ aims to explore unemployment complexities by inspecting its causes, consequences and appropriate solutions within Kerala. We try to contribute to a deeper understanding of unemployment dynamics by relevant data analysing and help to inform policy makers and authorities so that necessary actions could be taken to diminish the rate of unemployment.

OBJECIVES

The main objectives of the study are as follows:

1. To model and forecast rate of unemployment in rural and urban areas using seasonal ARIMA.
2. To model and forecast rate of unemployment in rural and urban areas using Holt-Winters.
3. To compare the forecast of the best model.

CHAPTER 2

LITERATURE REVIEW

1. Ullah et al.(2016) intended to explore the best forecasting model for forecasting unemployment among ARIMA, ARFIMA and exponential smoothing. The study analysed unemployment using various time series techniques. Measuring of long & short run relationship with population growth and participation rate of labour force and crop production was done. The study also investigated the causality between unemployment and the other variables. For the analysis, time series data from 1965 to 2014 was collected from Pakistan Economic Survey. The forecasting performance of three models was evaluated by using the forecast accuracy criterion such MAE, MAPE, RMSE and Theil's U statistics. On the basis of forecast accuracy criterion, Double Exponential Smoothing model was chosen as a best forecasted model for unemployment rate. In order to check the stationarity in the variables, ADF and PP test was used. Initially the variables were non stationary and became stationary at first differencing. Johnson cointegration and VECM results showed that there were long & short run cointegration relationship between unemployment rate and the other variables. The bi-directional causality running from crop production towards population growth was estimated using Granger Causality test.

2. Khrais & Al-Wadi (2016) examined the relationship between unemployment and Gross Domestic Production growth and was done in MENA countries. The results showed that the impact values considered by GDP on the Unemployment in all the countries being involved. The significance level of (F) was greater than ($\alpha = 0.05$) suggesting that there was no significant impact observed for GDP (annual) representing all the countries being involved in the study on Unemployment in all the countries. (- 0.009) was the impact value and is considered to be very small. This value suggests that there can be other factors too which affects unemployment apart from GDP.

3. Urrutia (2017) intended to formulate a mathematical model in order to forecast and estimate unemployment rate in the Philippines. Also, among the

considered variables namely Labor Force Rate, Population, Inflation Rate, Gross Domestic Product, and Gross National Income the factors which can predict the unemployment is to be found out. Using Pairwise Granger-causality test and Johansen Cointegration Test, Granger-causal relationship and integration among the dependent and independent variables are also examined. The model for forecasting the unemployment rate is SARIMA (6, 1, 5) × (0, 1, 1)4 with a coefficient of determination of 0.79 and is determined using the Box-Jenkins method. The actual values the predicted values were 99 percent similar and were 72 percent closely relative to the forecasted values. Labor Force Rate and Population were the significant factors of unemployment rate according to the results of the regression analysis. Among the independent variables, Population, GDP, and GNI showed that there is a granger-causal relationship with unemployment. There was a minimum of four cointegrating relations between the dependent and independent variables.

4. Katris (2020) analysed time series and machine learning models for prediction of unemployment in various countries (Med, Baltic, Balkan, Nordic, Benelux) for different forecasting horizons. When long memory exists in a time series FARIMA was a suitable model and has been applied successfully for predicting unemployment. The issue of heteroskedasticity was solved as he explored whether FARIMA models with GARCH errors achieve more accurate results. Models with non-normal errors were considered to improve forecasting accuracy. However, the non-linearity of the data could not be taken into account by the above models and due to this fact, to forecast unemployment rates he employed three machine learning techniques i.e. fully connected feed forward neural networks, support vector regression and multivariate adaptive regression splines. ARIMA and Holt-Winters were considered as benchmark models. Finally, the effects of different forecasting horizons and different geographic locations in terms of accuracy of forecasting of the models were explored.

5. Thomakos & Kyriazi (2020) reported on time series modelling and forecasting using several US. The Hendry and Doornik automatic time series PC-Give (Auto Metrics) methodology was applied to the macroeconomic series, US. real GDP and the unemployment rate. The LEI are a statistically important input to real GDP. The forecasting ability of best univariate and best bivariate models over 60- and 120-period rolling windows were tested and considerable forecast error reductions were

reported. The ADA-AR and the ADL, generated the smallest root-mean-square errors and lowest mean absolute errors respectively. Their results were greatly supportive of the importance for modelling and forecasting of the suggested input variables. Necessary improvements were also made over all traditional benchmarks.

6. Chakraborty et al.(2020) based on linear and nonlinear models suggested an integrated approach which can forecast the unemployment rates more correctly. The model proposed of the unemployment rate helps to improve their forecasts by reflecting the asymmetry of unemployment rate. From various countries, namely, Canada, Germany, Japan, Netherlands, New Zealand, Sweden, and Switzerland, the model's applications were shown. This was done using seven unemployment rate data sets. The results of computational tests were very promising on comparing with other conventional methods. The results for asymptotic stationarity of the suggested hybrid approach using Markov chains and nonlinear time series analysis techniques were given in this paper. This ensures that the proposed model could not show 'explosive' behaviour.

7. Adesegun et al.(2020) evaluated the trend, model and forecast for the future rate of unemployment in Nigeria. The statistical methodology used in this research work was univariate time series analysis using autoregressive integrated moving averages (ARIMA). From the time plot it was seen that rate of unemployment in Nigeria was not stable over the past years. In order to find out if the data is stationary, ADF test was done. The data was transformed through differencing. Then, KPSS test was used to confirm the stationarity of the process ($P > 0.05$). Various ARIMA (p, d, q) were observed such as ARIMA (1,1,1), ARIMA (2,1,1), ARIMA (2,1,2), ARIMA (2,2,1) and ARIMA (2,2,2) with AICs and Log-likelihood (107.52, −50.76), (109.22, −50.61), (111.08, −50.54), (110.08, −51.04) and (111.36, −50.68) respectively. Due to the least AIC and highest log likelihood, ARIMA (1,1,1) was selected. Both Shapiro-Wilk test and Box test was performed to confirm the fitness of the model ($P > 0.05$) for the process. Forecast for 5 years was then made for the process. In conclusion, the model obtained in this research can be used for making inference, monitoring and controlling the unemployment rates in Nigeria. Necessary solutions such as industrialization, diversification in the economy, investment in agriculture, modification in education curriculum have been recognised. The analysis is completely done in R package.

8. Sharma & Soni (2021) forecasted the unemployment rate of youth in India for five years from 2019 to 2023. The data on the youth unemployment rate in India from 1991 till the end of 2018 was taken. The frequency of the time was taken to be annual. Various models were fitted onto this time series individually after the analysis of historical data available for the purpose of predicting the future unemployment rate. They have compared various exponential smoothing models and ARIMA model onto the time series data. This made them conclude that the ARIMA model performs best among all the other models on the basis of accuracy of forecasting.

9. Nkoane & Seeletse (2021) learned to build the time series model and forecast the unemployment rate in South Africa. It was done in the presence of the unexpected events or contamination of data using robust estimators. Robust estimators deal with outliers while identifying the orders, for estimating the parameters of the time series models. Time series data are usually contaminated with the anomalies and the standard methods of estimating the parameters such as ML, LS and MM are sensitive to anomalies. The unemployment time series data taken is from January 2010 to December 2020 and is quarterly. Though, outliers are identified are not removed. An ARIMA (1, 1, 1) model is found to be the best model for the unemployment series. The standard methods, such as the RMSE, MAPE, and MAE are used to measure the accuracy of the forecast. They concluded that the forecasting results for the unemployment rate show that better performance was made by the robust estimators in the presence of outliers.

10. Kramarova et al.(2022) intended to develop a complete analysis of unemployment along with its development in Slovakia, during the period of Covid-19 pandemic. The aim of the study was to determine to which extent the anti-pandemic measures have affected the labour market of Slovak. The study also wanted to recognise the most affected groups of unemployed among all the people. Methods taken were to analyze the impact of the anti-pandemic measures on the unemployment condition in Slovakia. They applied the method where the statistically created imaginary state of the absence of the covid pandemic was compared along with the real state as the consequence of the pandemic. The imaginary state was modelled by the one-dimensional time series model with a linear trend and seasonability. Findings done were that the analysis results identify the population groups, who were typically affected by the pandemic along with the quantification of the impact on

unemployment. These findings can be further used in to create targeted measures aimed at supporting individuals who are unemployed at the sustainability of already existing jobs and in other governmental economic and social decisions.

CHAPTER 3

MATERIALS AND METHODS

3.1 DATA DESCRIPTION

The dataset contains data on unemployment rate in India from Jan 2016 to Feb 2023. The data is given monthly wise and includes the attributes such as Date and Estimated Unemployment Rate (%). It has been collected from Kaggle which gives the data of both rural and urban areas.

3.2 TOOLS FOR ANALYSIS

i)Seasonal ARIMA (Auto Regressive Integrated Moving Average) model: SARIMA is an extended version of ARIMA which consists of seasonal components as well to account for the periodic patterns in the data. It comprises of seasonal autoregressive (SAR) and seasonal moving average (SMA) terms. A seasonal ARIMA model is denoted by ARIMA (p, d, q) × (P, D, Q). SARIMA is usually used if the data is monthly or if it exhibits seasonality.

ii)Holt-Winters model: This method is suitable for time series data having trend and seasonal components. Three types of smoothing techniques are involved: level smoothing, trend smoothing, and seasonal smoothing

The process of fitting the model and forecasting is done using Python.

3.3 TOOLS FOR COMPARISON

i) Mean Squared Error (MSE): MSE is a statistical tool used in regression problems. MSE is calculated by taking the average of the squared differences between the actual target values and the predicted values produced by the model. If the MSE value is less, it indicates that the model forecasting has come closer to the actual values, whereas a higher MSE value indicates larger errors between predicted and actual values.

ii) Root Mean Squared Error (RMSE): RMSE is another tool which is used in statistics and machine learning to identify the regression model performances. It is calculated from Mean Squared Error (MSE). For RMSE computation, we first calculate the MSE value and the square root of the result is considered to be the RMSE value.

iii) Akaike Information Criterion (AIC): AIC means Akaike Information Criterion which is used for choosing of model among a set of other models. It compares the goodness of fit of different models while balancing for the complexity and model fit of the model. AIC is commonly used in different fields such as statistics, machine learning etc, so that the best model can be chosen**.**

The chosen data sample is as follows:

Rural region data sample,

DATE	ESTIMATED UNEMPLOYMENT RATE(%)
31-01-2016	20.60
29-02-2016	23.68
31-03-2016	26.37
30-04-2016	17.39
31-05-2016	13.10

Urban region data sample,

DATE	ESTIMATED UNEMPLOYMENT RATE(%)
31-01-2016	15.87
29-02-2016	23.35
31-03-2016	18.61
30-04-2016	19.55
31-05-2016	10.53

CHAPTER 5

TIME SERIES

A sequence of data points or observations which is collected or recorded successively and equally spaced intervals of time is known as Time Series. Time series analysis is a statistical technique which is used for analysing and interpreting data points collected. Various fields such as economics, engineering, finance, environmental science and signal processing use time series to obtain necessary insights. Its main objective is to recognise the underlying patterns and trends which exists in the data over time. A particular phenomenon, variable, or system evolves and changes over time can be analysed and identified using Time series.

4.1 COMPONENTS OF TIME SERIES

The main components of time series are as follows:

1) Trend: The long-term movement in the data is usually represented by the trend component. It can be increasing, decreasing or can stay relatively constant over time. Trends can be linear or nonlinear which means they show a straight line or show more complex patterns respectively. In order to identify the existing patterns and forecasting, it is necessary that the trend component is to be identified.
2) Seasonality: The predictable fluctuations in the data that occur at fixed intervals are referred by seasonality. They can be daily, weekly, monthly or yearly. Changes in weather, holidays and periodic factors are often associated by seasonal patterns. It is crucial to account for seasonality so that accurate short-term forecasts can be made. It also helps in identifying the cyclic nature of certain phenomena.
3) Cyclicality: Cyclicality includes non-seasonal fluctuations in the data which repeats at irregular intervals and is dissimilar to seasonality of data. These cycles are associated with longer-term patterns which may have varying time durations and amplitudes. It is challenging to identify cyclical patterns because they do not follow fixed time frames like seasonal patterns. However, it is important to analyse and understand cyclicality for identifying economic trends and decision making.

4) Irregular or Random Component: the residual variability in the data that cannot be explained by trend, seasonality, or cyclicality is represented by the random component. The random fluctuations, noise or unpredictability characteristic in the data is reflected by this component. The factors which contribute are measurement errors, outliers, or unforeseen events. The random component is essential for accurate modelling and predicting, as it can impact the reliability of predictions and insights which are obtained from the time series data.

4.2 MATHEMATICAL MODELS IN TIME SERIES

1) Additive model: According to the additive model, the decomposition of time series is done assuming that the effects of the components are additive. The components and level of the time series are assumed to be independent in additive model. Mathematically, it can be written as:

$$Y_t = T_t + S_t + C_t + I_t$$

Where:

Y_t - the observed value of the time series value

T_t - the trend component

S_t - the seasonal component

C_t – Cyclic variations

I_t - residual or random component

2) Multiplicative model: According to the additive model, the decomposition of time series is done assuming that the effects of the components are dependent of each other. Mathematically, it can be written as:

$$Y_t = T_t \times S_t \times C_t \times I_t$$

4.3 FORECASTING

Forecasting the future values is a significant problem in many fields such as economic, sales forecasting, stock control etc. Forecasting time series data involves predicting future values based on historical data. Necessary conclusions and insights can be obtained from the past data and thus future can be predicted. Various techniques and models can be used for forecasting of time series, depending on the data characteristics. Depending on the specific objectives of the analysis, we could choose which procedure and model should be adopted for forecasting. Some of the ways for the forecasting of time series data are:

1) Simple Moving Average (SMA): It forecasts the future values by taking the average of the past observations over a fixed range of time. This method is simple to implement and can help smooth out random fluctuations in the data. However, no trends or seasonality in the data are captured by SMA.

2) Exponential Smoothing methods: Exponential smoothing is used for forecasting of future values where exponentially declining weights are assigned to the past observations. On the basis of the smoothing parameter chosen, it is can capture trends and seasonality in the data.

Types of Exponential smoothing methods are:

i) Simple Exponential Smoothing (SES)
ii) Double Exponential Smoothing (Holt's method)
iii) Triple Exponential Smoothing (Holt-Winters method)

Holt-Winters Forecasting: This method is suitable for time series data having trend and seasonal components. Three types of smoothing techniques are involved: level smoothing, trend smoothing, and seasonal smoothing.

i) Level smoothening: The level smoothing component is often denoted as L_t. It signifies the smoothed level of the time series at t time.
The level smoothing equation is given by:

$$L_t = \alpha \times Y_t + (1- \alpha) \times (L_{t-1} + T_{t-1})$$

Where Y_t - the observed value of the time series value
T_{t-1} - the trend component from previous time period
L_{t-1} - smoothed level from the previous time period
α - level smoothing parameter and lies between 0 and 1

ii) Trend Smoothing: The trend smoothing component is denoted as T_t, the smoothed trend of the time series at time t is represented by the trend smoothing. Long-term changes in the data are identified by trend smoothing. The trend smoothing equation is given by:

$$T_t = \beta \times (L_t - L_{t-1}) + (1-\beta) \times T_{t-1}$$

Where L_t and L_{t-1} are the smoothed levels for the present and previous time periods respectively

T_{t-1} is the smoothed trend from the previous time period

β is the trend smoothing parameter ranging from 0 to 1

iii) Seasonal Smoothing: S_t represents the seasonal smoothing component. Seasonal variation is represented by this smoothing. Repetitively occurring cycles are identified. The equation for seasonal smoothing is given by,

$$S_t = \gamma \times (Y_t - L_t) + (1-\gamma) \times S_{t-m}$$

Where Y_t - the observed value of the time series value

L_t - smoothed levels for the present period

S_{t-m} - is the smoothed seasonal component from the same season in the previous year.

4.4 Time Series Modelling

4.4.1 Basic Definitions

i) Autoregressive (AR) processes

To describe the behaviour of a variable over time, a stochastic model is used in time series analysis which is known as Autoregressive (AR) process. It's considered to be a type of linear regression model. A linear combination of the past values and a random error term is modelled as the present value of the variable. Mathematically, an AR(p) process of order p can be represented as:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots .. + p_1 Z_{t-p} + \varepsilon_t$$

Where $\{\varepsilon_t\}$ – purely random process having mean of 0 and variance σ^2

ii) Moving Average (MA) processes

In order to explain the behaviour of a variable with time, a different type of stochastic model is used which is the Moving Average (MA) process. AR processes model the present value of a variable as a linear combination of its previous values whereas MA processes model the present value as a linear combination of previous error terms. Mathematically, it can be represented as:

$X_t = \mu + \varepsilon_t + \theta_1 \times \varepsilon_{t-1} + \theta_2 \times \varepsilon_{t-2} + \ldots + \theta_q \times \varepsilon_{t-q}$

Where:

X_t - value of the time series at time t.

μ - constant term.

ε_t - white noise error term at time t representing the random fluctuations or shocks in the data.

$\theta_1, \theta_2 \ldots$..are the moving average parameters

iii) SARIMA (Seasonal Autoregressive Integrated Moving Average)

SARIMA is an extended version of ARIMA which consists of seasonal components as well to account for the periodic patterns in the data. It comprises of seasonal autoregressive (SAR) and seasonal moving average (SMA) terms. A seasonal ARIMA model is denoted by ARIMA (p, d, q) × (P, D, Q). SARIMA is usually used if the data is monthly or if it exhibits seasonality.

iv) Stationarity

If the mean, variance, and autocorrelation of a time series data do not change as the time passes then it is said to be a stationary time series. For accurate analysis and forecasting it is significant as most of the time series models and statistical techniques need the data to be stationary. Once the time series is stationary, it is easier to model and analyse the data. Various time series models such as ARIMA and SARIMA work with stationary data. Thus, the data has to be converted to be converted to stationary if it is not. This can be done by using various tests, such as Augmented Dickey-Fuller (ADF) test.

4.4.2 Box-Jenkins Modelling Procedure

It involves model identification where the model is recognised. Then the parameters are estimated. Diagnostic checking can be done to estimate if the model is adequate for the problem. If yes, the model can be used to find the forecasted values**.** In diagnostic checking, the residual series is carefully analysed. One can construct a histogram of standardised residuals and compare it with the standard normal distribution and see if the errors are normally distributed. Using a residual plot, we can check if the random shocks have a mean of zero and constant variance.

4.4.3 Akaike Information Criterion (AIC)

AIC means Akaike Information Criterion which is used for choosing of model among a set of other models. It compares the goodness of fit of different models while balancing for the complexity and model fit of the model. AIC is commonly used in different fields such as statistics, machine learning etc, so that the best model can be chosen**.**

4.4.4 Statistical tests

1) Augmented Dickey-Fuller (ADF) test

This test comes under the category of 'Unit Root Test'. It is one of the best methods of checking stationarity of a time series data. It is a commonly used test when it comes to analysing the stationarity of a series.

It tests the following cases:

H_0: The time series is non-stationary.

H_1: The time series is stationary.

If the p-value is less than a particular value (eg.,0.05), then the null hypothesis is rejected an concludes that the time series is stationary.

2) The KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test

It is another method to ensure stationarity of the time series data. It tests the following cases:

H_0: The time series is stationary.

H_1: The time series is non-stationary.

If the p-value is greater than a particular value (eg.,0.05), then the null hypothesis is rejected an concludes that the time series is non-stationary.

CHAPTER 5

DATA ANALYSIS AND FORECASTING

5.1 TIME SERIES MODELLING

This chapter presents the time series modelling of unemployment rate in rural and urban areas of Kerala from 2016 to 2023. The process of model fitting was done by using Python programming language.

5.1.1 Modelling and Forecasting of unemployment rate in Urban areas using SARIMA

Fig 4.1 depicts the time series plot of unemployment rate in rural areas. It is visible that there is small decrease in variance, seasonality and a downward trend in the data.

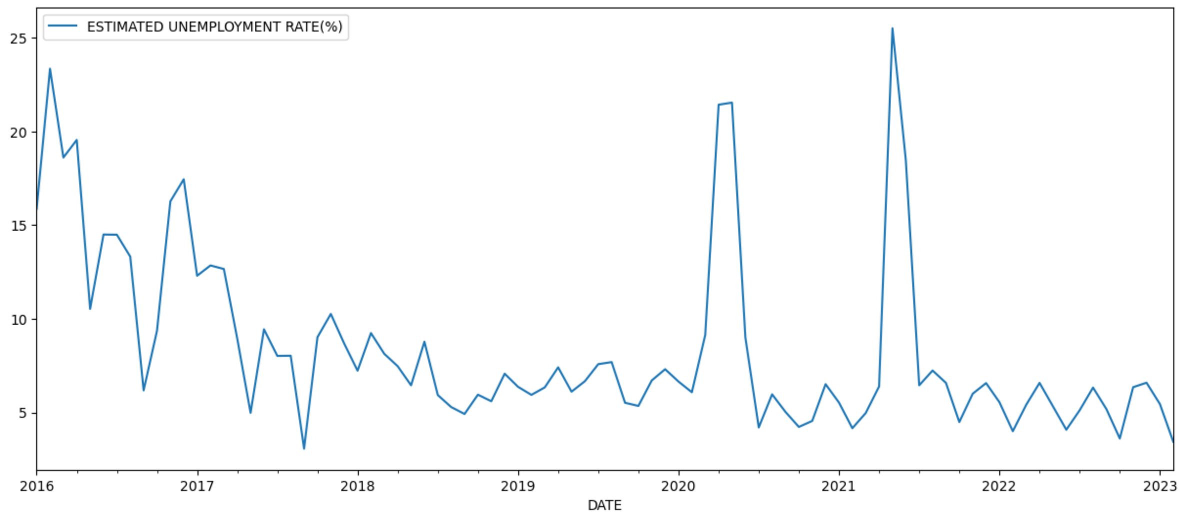

Fig 5.1: time series plot of unemployment rate from 2016 to 2023

5.1.1.1 Augmented Dicky Fuller (ADF) Test

Augmented Dickey Fuller test (ADF Test) is a common statistical test which is used to test whether a given Time series is stationary or not.

It is considered as one of the most commonly used tests in order to analyse the stationarity of a time series data.

We have obtained from the test that the p value is 1.2418621259212833e-07 which is less than 0.05. Thus, we reject the null hypothesis "Time series is non stationary", which means the time series is stationary.

5.1.1.2 Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test

KPSS test is another statistical test of checking the stationarity of a time series data. We have obtained that the p value is 0.1 which is greater than 0.05. Thus, we conclude that the time series is stationary.

Table 5.1 shows the possible time series models along with their corresponding values

Sl.no	**Model** **ARIMA (p, d, q) × ARIMA (P, D, Q)**	**AIC Values**
1	ARIMA (0, 0, 0) x (0, 0, 0)	637.7481970474067
2	ARIMA (0, 0, 0) × (0, 0, 1)	591.7554108751526
3	ARIMA (0, 0, 0) × (0, 1, 0)	460.17389622216615
4	ARIMA (0, 0, 0) × (0, 1, 1)	459.00506761404887
5	ARIMA (1, 1, 1) × (0, 0, 1)	467.14952003163745
6	ARIMA (1, 1, 1) × (0, 1, 0)	432.19319741851814
7	**ARIMA (1, 1, 1) × (0, 1, 1)**	**422.92313384411636**
8	ARIMA (1, 1, 1) × (1, 0, 0)	467.18200982900083
9	ARIMA (1, 1, 1) × (1, 0, 1)	469.0340171273163
10	ARIMA (1, 1, 1) × (1, 1, 0)	426.4272433136628
11	ARIMA (1, 1, 1) × (1, 1, 1)	424.4397494072726

Table 5.1

According to Akaike Information Criterion (AIC), ARIMA (1,1,1) × (0,1,1) model is considered as the most appropriate one.

Fig 5.2 shows the presence of trend and seasonality in the time series data

Thus, SARIMA (Seasonal Auto Regressive Integrated Moving Average) can be used for forecasting of future values.

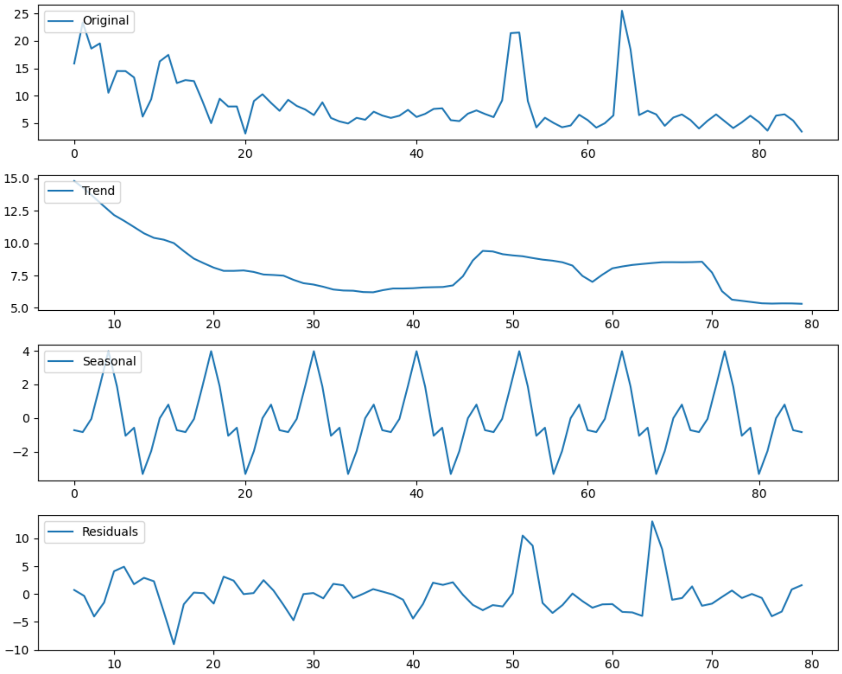

Fig 5.2: seasonal decomposition plot

5.1.1.3 DIAGNOSTIC CHECKING

Diagnostic checking is the process of evaluating the capability of a fitted model by examining the residuals. The main aim of diagnostic checking is to ensure that the assumptions of the model are met and the residuals exhibit certain expedient properties. The below figures show that the data is normally distributed. Standard residuals for 'E' show how much outliers are present in the dataset and the Q-Q plot shows the normality of the data.

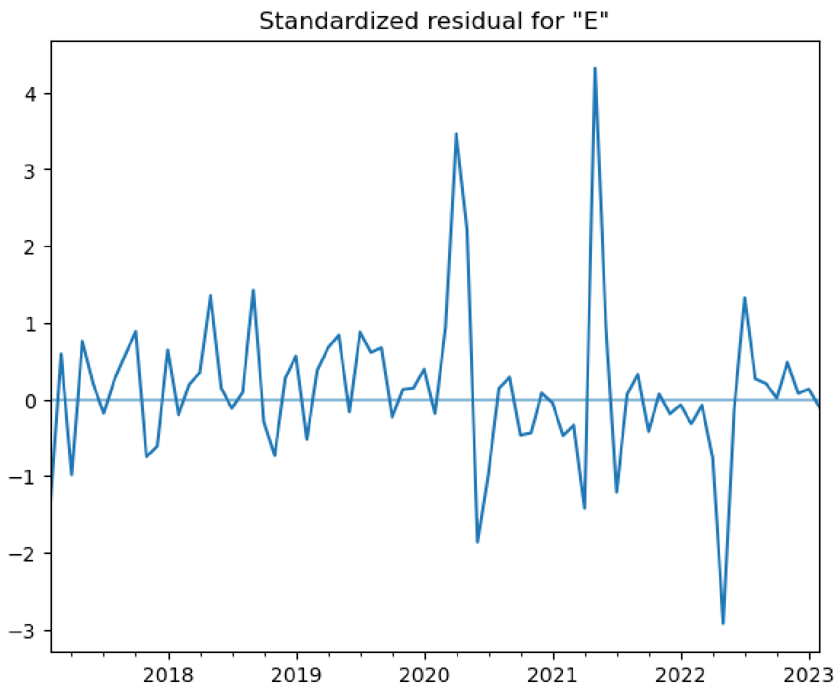

Fig 5.3: standardized residual for "E"

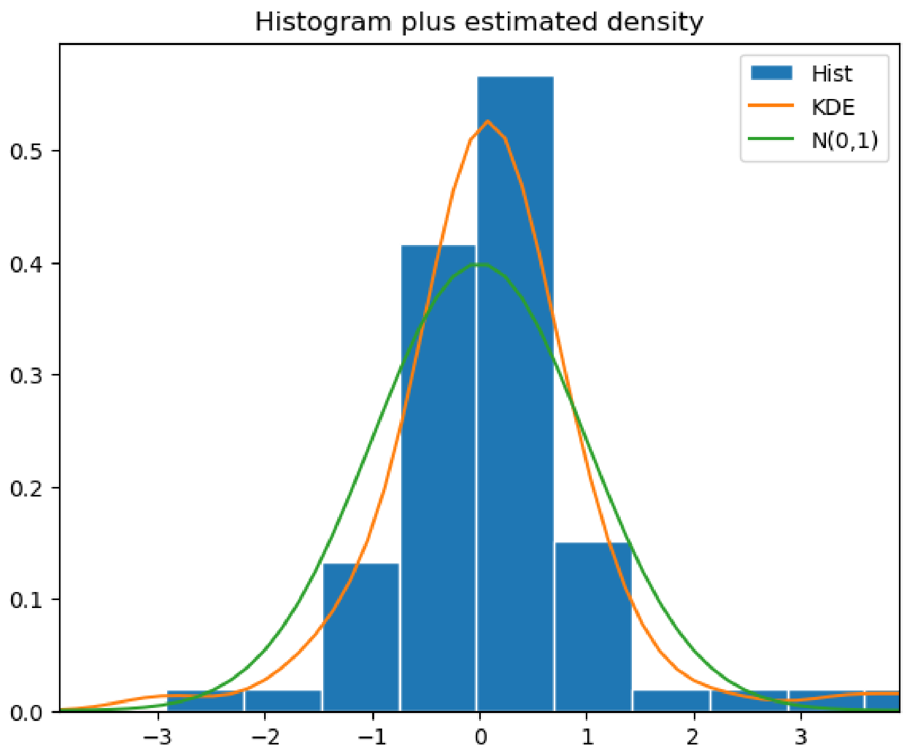

Fig 5.4: histogram plus estimated density of SARIMA model

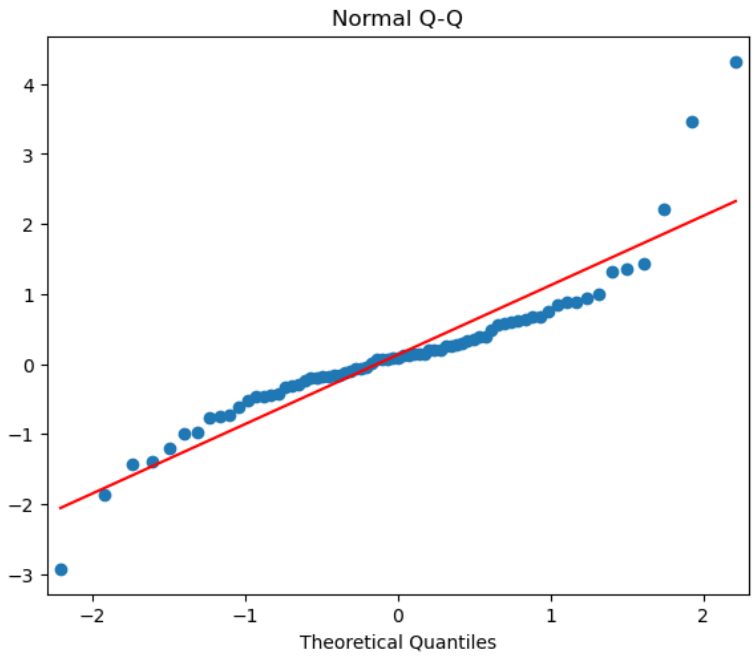

Fig 5.5: Normal Q-Q plot of noise residuals of the SARIMA model

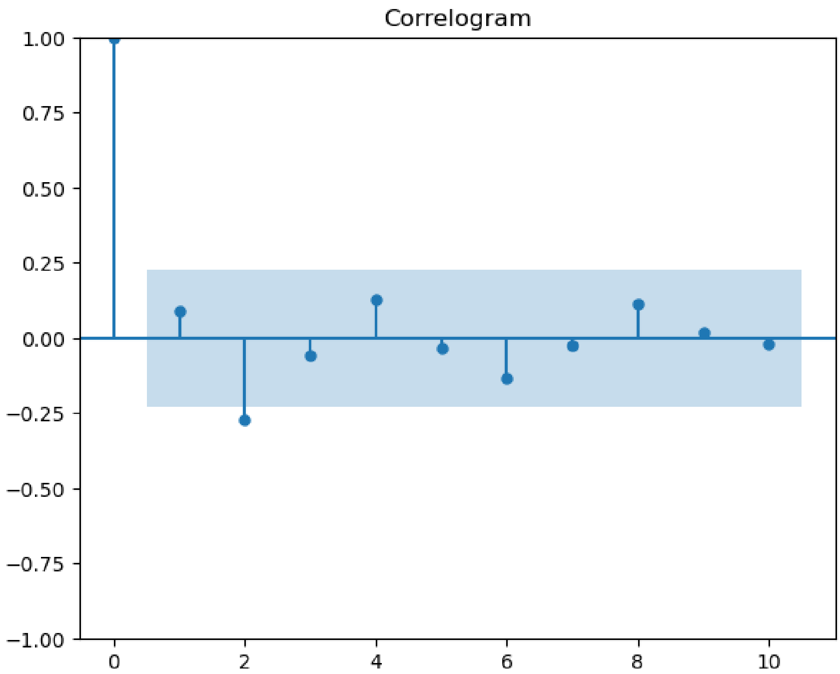

Fig 5.6: the Correlogram

The In-sample forecast is obtained as follows:

Date	Actual values	Predicted values
2016-02-29	23.35	15.869952
2016-03-31	18.61	23.349931
2016-04-30	19.55	18.609994
2016-05-31	10.53	19.549991
2016-06-30	14.50	10.530050
2016-07-31	14.49	14.500002
2016-08-31	13.32	14.490002
…	…	…
2022-07-31	5.12	-0.192494
2022-08-31	6.33	5.255743
2022-09-30	5.18	4.357430
2022-10-31	3.61	3.542428
2022-11-30	6.35	4.410977
2022-12-31	6.59	6.278565
2023-01-31	5.45	4.912677
2023-02-28	3.44	3.861898

Table 5.2 shows the In-sample forecast using SARIMA model

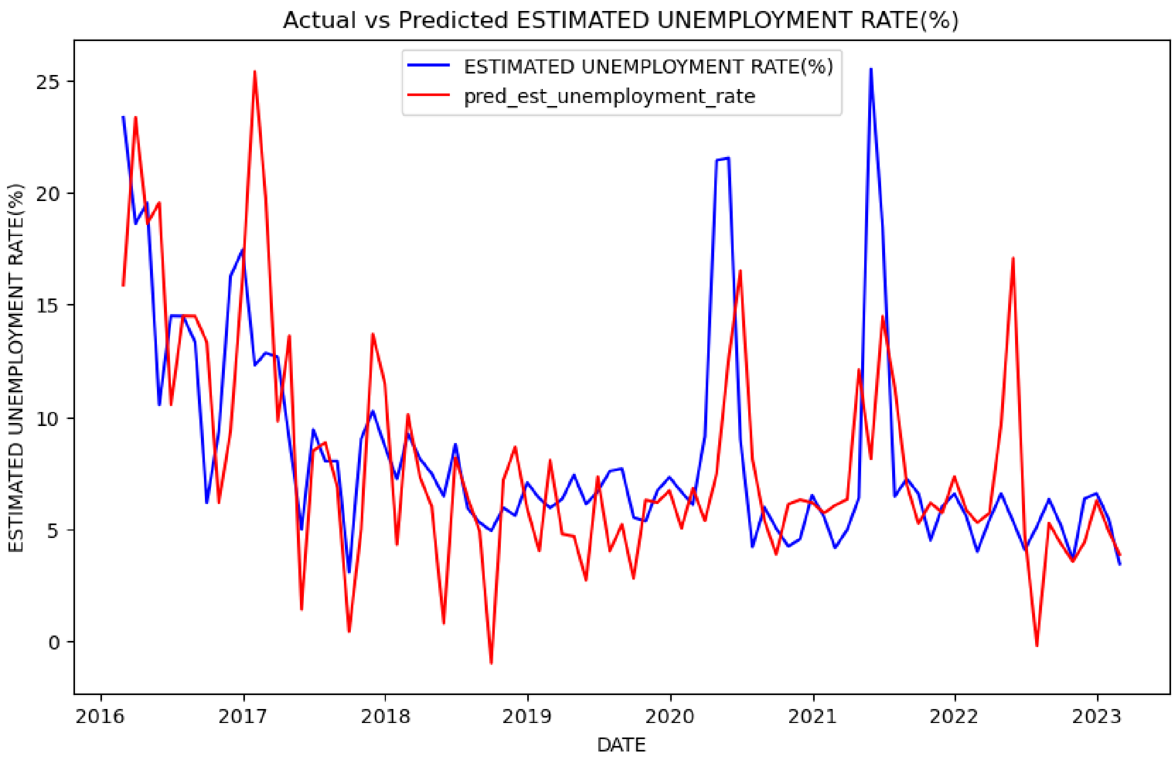

Fig 5.7: fitted verses actual values using the SARIMA model

5.1.1.4 Forecasting of Estimated Unemployment Rate (%) using SARIMA model

The LCL and UCL values are given in the following table,

LCL	UCL
-3.311226	12.442209
-2.151926	16.131910
1.220228	20.472811
-2.762562	16.985953
-6.076872	13.993189
-5.207270	15.111059
-6.493955	14.038842
-7.835644	12.894880
-5.965216	14.954176
-5.476344	15.626875
-6.662196	14.621711
-8.282198	13.179633
-8.352334	15.596177
-6.461074	18.601831
-2.919784	22.797651
-6.884490	19.304787
-10.230210	16.349497
-9.410402	17.520206
-10.754178	16.506608
-12.155798	15.423345
-10.346197	17.543854
-9.918165	18.277586
-11.164478	17.332980
-12.844259	15.950729
-12.871256	18.324160
-11.026699	21.376528
-7.546725	25.633671
-11.574065	22.203445
-14.981414	19.309785
-14.222022	20.540909

Table 5.2

The forecasted values along with the corresponding plot are given in table 5.3 and figure 5.8 respectively.

Date	Forecast values
2023-03-31	4.565491
2023-04-30	6.989992
2023-05-31	10.846519
2023-06-30	7.111696
2023-07-31	3.958159
2023-08-31	4.951894
2023-09-30	3.772444
2023-10-31	2.529618
2023-11-30	4.494480
2023-12-31	5.075266
2024-01-31	3.979758
2024-02-29	2.448717
2024-03-31	3.621922
2024-04-30	6.070379
2024-05-31	9.938934
2024-06-30	6.210149
2024-07-31	3.059644
2024-08-31	4.054902
2024-09-30	2.876215
2024-10-31	1.633774
2024-11-30	3.598828
2024-12-31	4.179710
2025-01-31	3.084251

2025-02-28	1.553235
2025-03-31	2.726452
2025-04-30	5.174915
2025-05-31	9.043473
2025-06-30	5.314690
2025-07-31	2.164185
2025-08-31	3.159444

Table 5.3: the SARIMA forecast

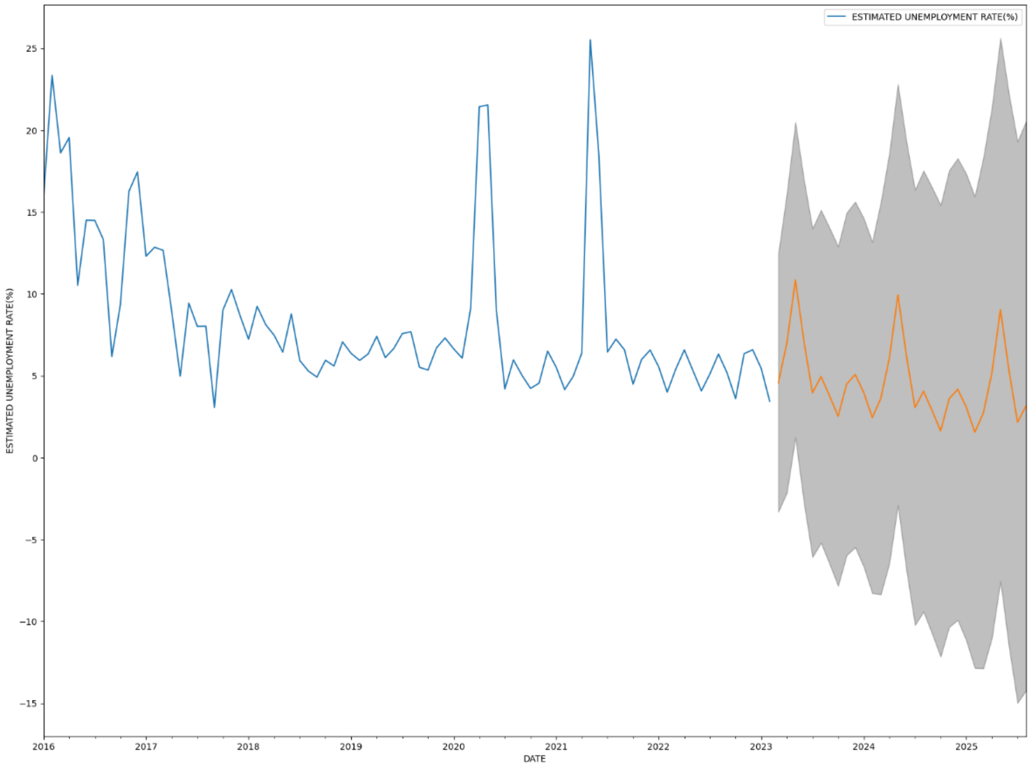

Fig 5.8: plot of the forecasted values using SARIMA model

5.1.2 Modelling and Forecasting of unemployment rate in Rural areas using SARIMA

Fig 5.9 depicts the time series plot of unemployment rate in rural areas. It is visible that there is small decrease in variance, seasonality and an upward trend in the data.

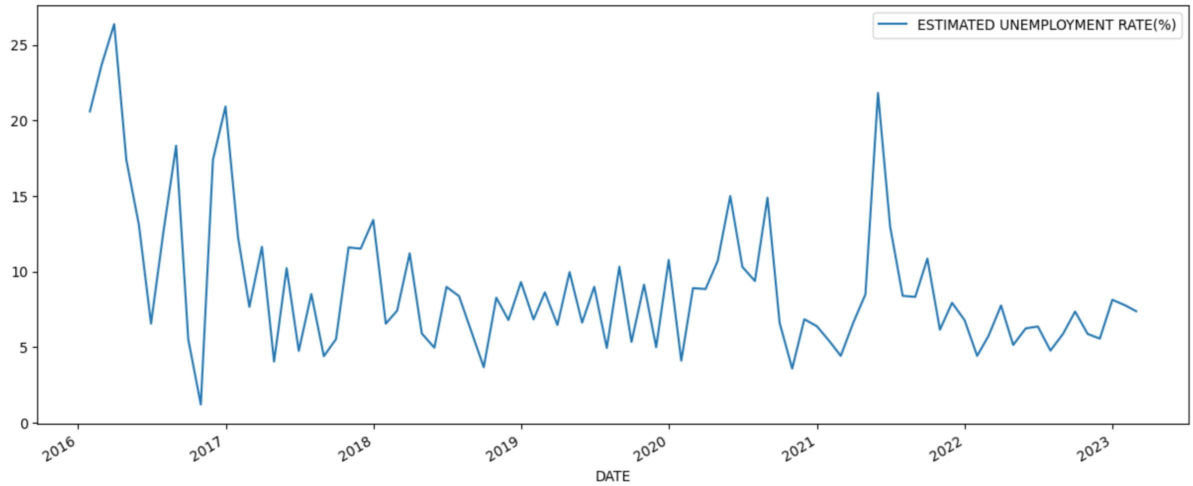

Fig 5.9: the time series plot of unemployment rate from 2016 to 2023

Using ADF test, we have obtained that the p value is 5.749491397606646e-07 which is less than 0.05. Thus, we reject the null hypothesis "Time series is non stationary", which means the time series is stationary.

The best model can be achieved from the AIC values of the model. The model having the least AIC value is considered to be the best one.

Sl.no	**MODEL** **ARIMA (p, d, q) × ARIMA (P, D, Q)**	**AIC values**
1	ARIMA (0, 0, 0) × (0, 0, 0)	637.7388722661109
2	ARIMA (0, 0, 0) × (0, 0, 1)	590.4547760731231
3	ARIMA (0, 0, 0) × (0, 1, 0)	452.48829780476666
4	ARIMA (0, 0, 0) × (0, 1, 1)	453.731540659012
5	ARIMA (0, 0, 0) × (1, 0, 0)	545.0546475275783

6	ARIMA (1, 1, 1) × (0, 1, 0)	434.6258840766475
7	ARIMA (1, 1, 1) × (0, 1, 1)	432.13797057762645
8	ARIMA (1, 1, 1) × (1, 0, 0)	480.3071522820377
9	ARIMA (1, 1, 1) × (1, 0, 1)	481.35674751951996
10	ARIMA (1, 1, 1) × (1, 1, 0)	434.03083895198756
11	**ARIMA (1, 1, 1) × (1, 1, 1)**	**431.801302976515**

Table 5.4: time series models and AIC values

According to Akaike Information Criterion (AIC), ARIMA (1,1,1) × (1,1,1) model is considered as the most appropriate one.

The below figure helps us to identify the presence of trend and seasonality in the time series data. Thus, SARIMA (Seasonal Auto Regressive Integrated Moving Average) can be used for forecasting of future values.

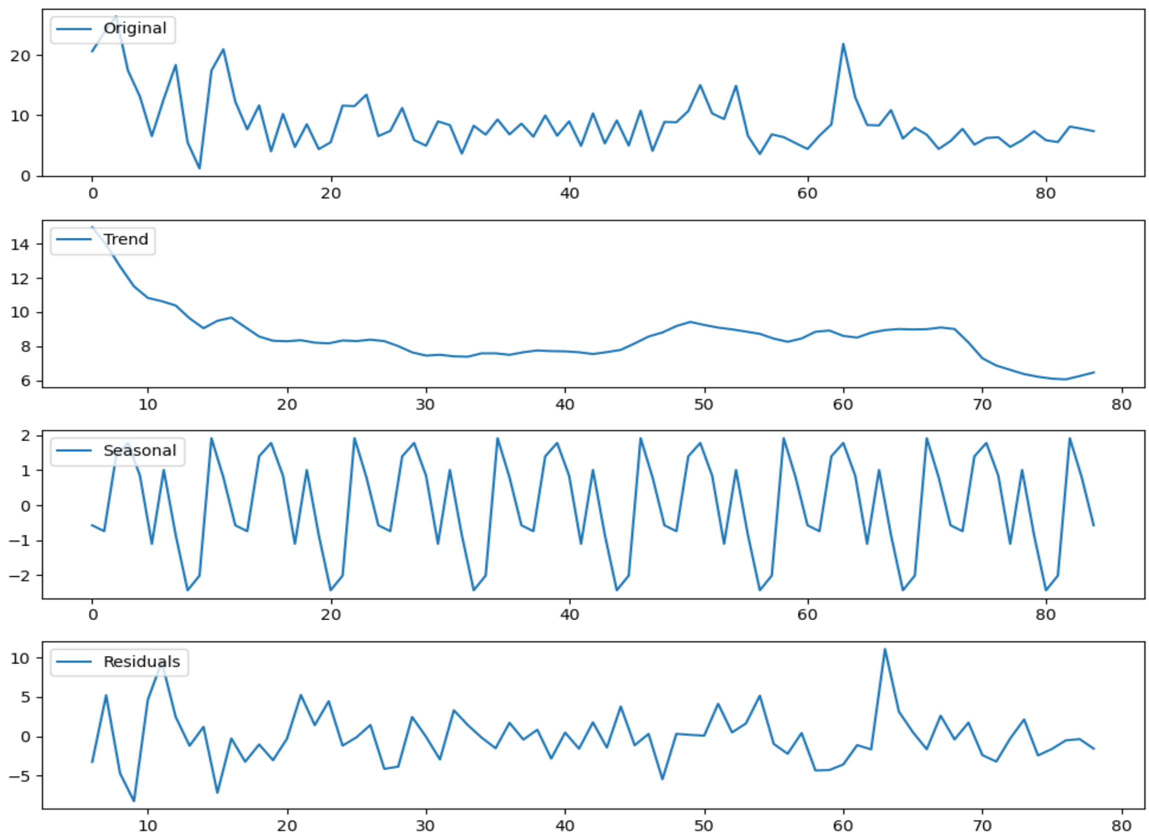

Fig 5.10 seasonal decomposition plot

From the diagnostic plot it is clear that, the data is normally distributed. Standard residuals for 'E' show how much outliers are present in the dataset. Normal Q-Q plot shows the normality of the data.

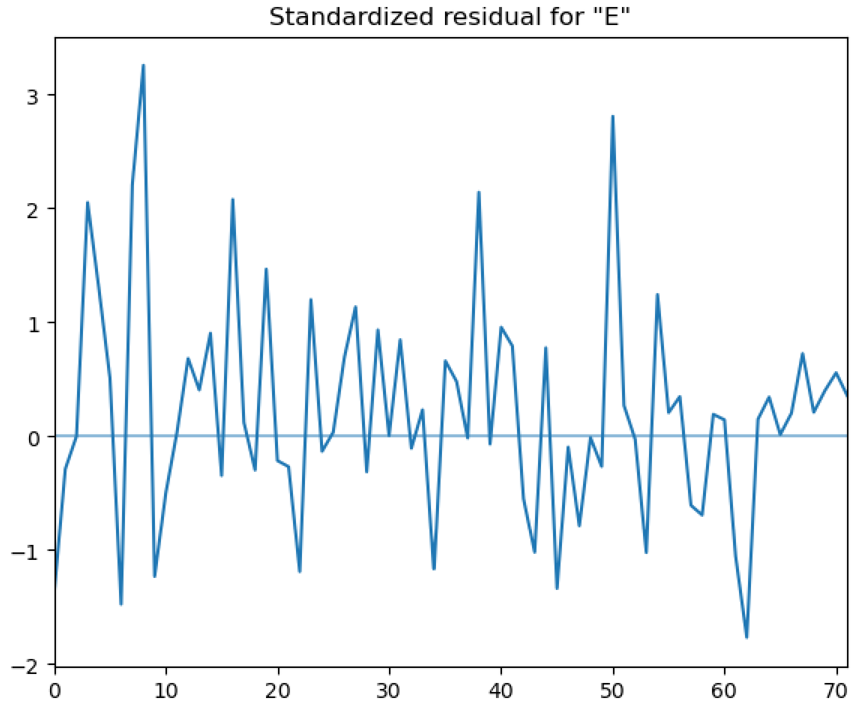

Fig 5.11 the standardized residual for "E"

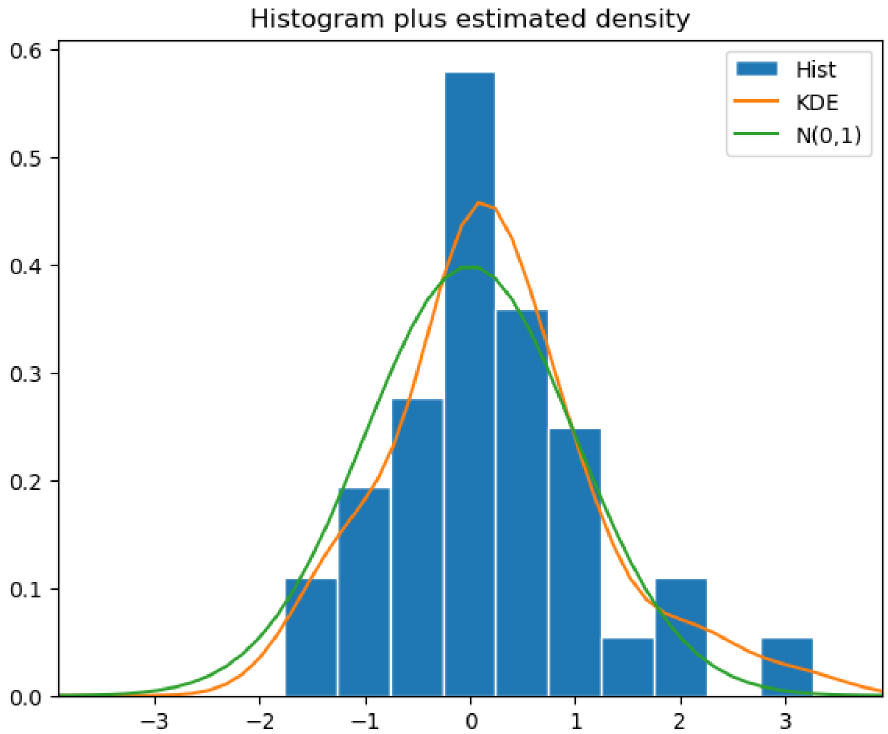

Fig 5.12 shows the histogram plus estimated density of SARIMA model

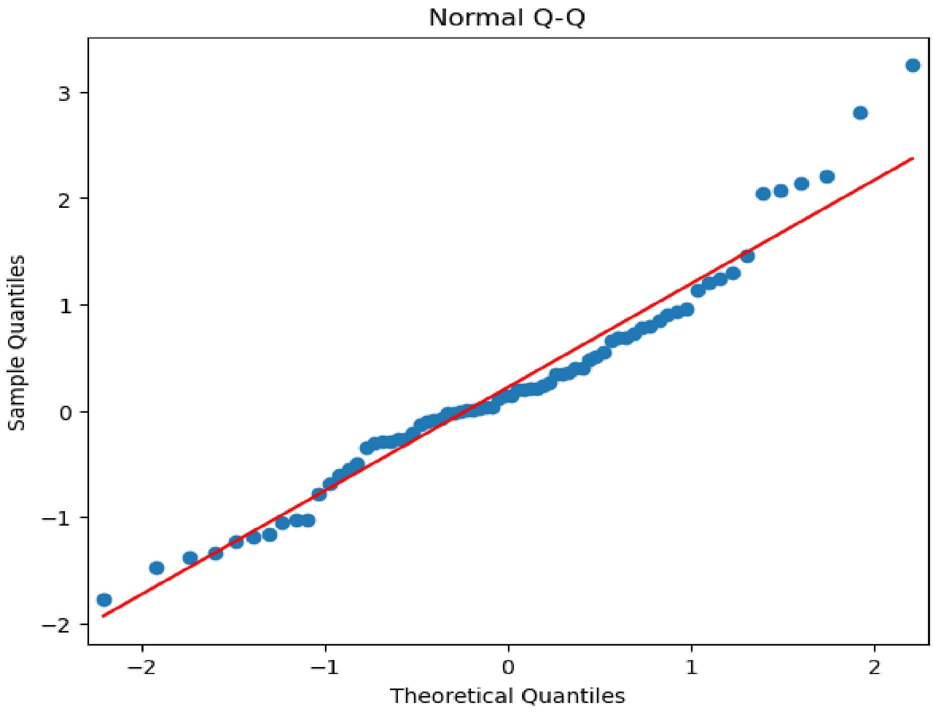

Fig 5.13: Normal Q-Q plot of noise residuals

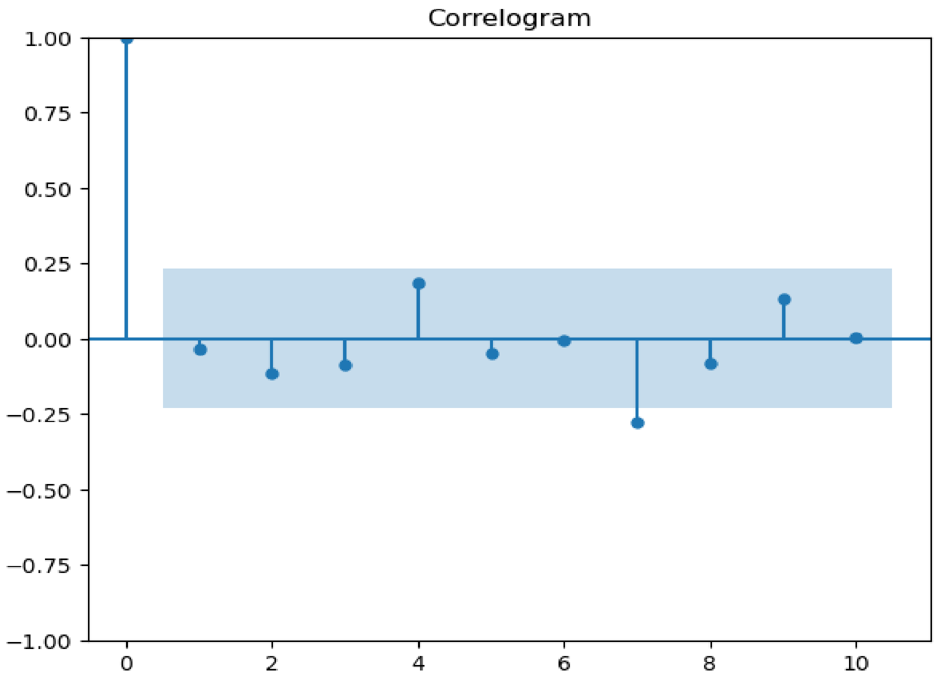

Fig 5.14 shows the Correlogram

The In-sample forecast using SARIMA model is tabled below,

Date	Actual values	Predicted values
2016-02-29	23.35	15.869952
2016-03-31	26.37	23.679936
2016-04-30	17.39	26.369953
2016-05-31	13.10	17.390088
2016-06-30	6.56	13.100071
2016-07-31	12.70	6.560091
…	…	…
2022-08-31	5.88	5.808897
2022-09-30	7.36	6.472772
2022-10-31	5.88	2.722942
2022-11-30	5.57	4.650306
2022-12-31	8.14	6.393769
2023-01-31	7.78	5.355854
2023-02-28	7.38	5.830042

Table 5.5: In-sample forecast

On plotting these values, we obtain the fitted values verses actual values graph where the x-axis shows the DATE and Estimated Unemployment rate (%) on the y axis.

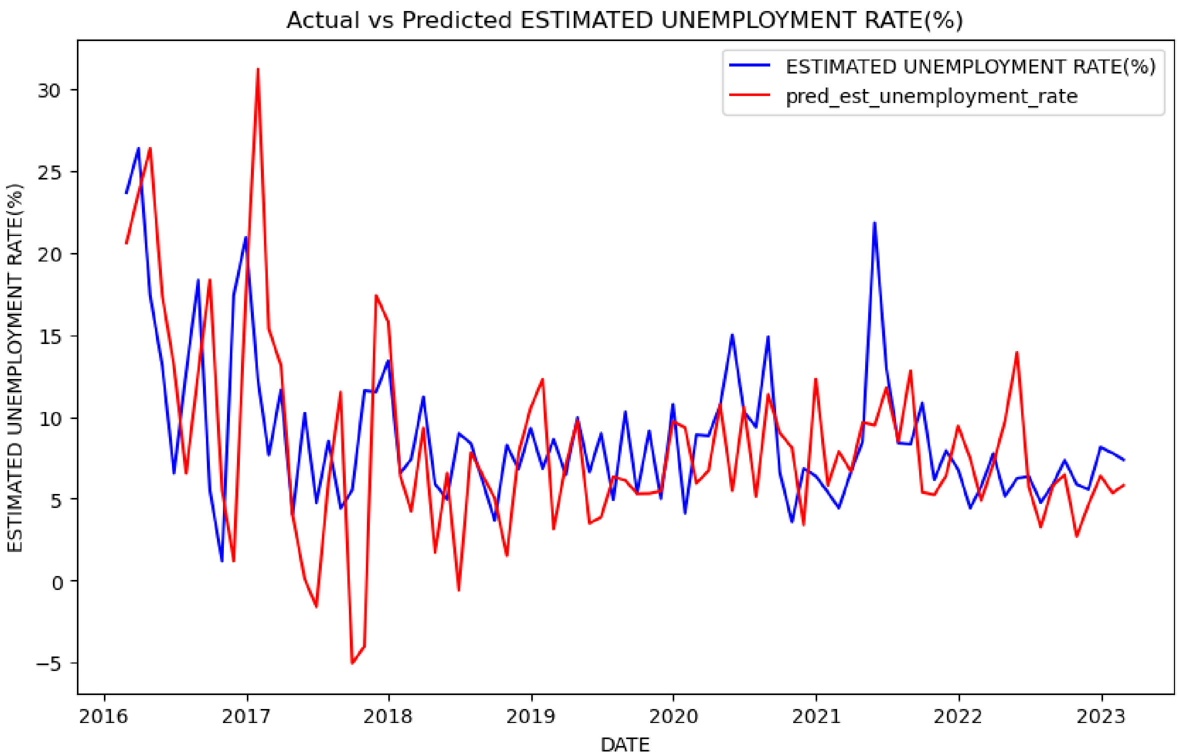

Fig 5.15: fitted verses actual values

5.1.2.1 Forecasting of Estimated Unemployment Rate (%) using SARIMA model

The LCL and UCL values are given by,

LCL	UCL
-1.131499	15.817149
-2.538563	15.785889
-3.165511	15.672511
-4.128157	15.068210
-6.622160	12.893593
-4.611931	15.210079
-4.391313	15.730323
-6.758503	13.658030
-7.251027	13.456751
-3.643428	17.353975

-4.182881	17.107562
-5.368066	16.209850
-7.085424	18.324687
-7.127394	19.378089
-8.458786	18.717125
-9.941171	17.805158
-12.897443	15.387415
-10.482752	18.325481
-10.902235	18.418836
-13.729833	16.095301
-14.342962	15.979165
-10.268670	20.547514
-10.928802	20.390095
-12.535513	19.296657
-13.325150	20.710791
-12.708572	22.284315
-14.361906	21.342893
-16.081378	20.269917
-19.249289	17.724420
-16.630532	20.952089

Table 5.6

The forecasted values along with the corresponding plot are given in table 5.7 and figure 5.17 respectively.

Date	**Forecast Values**
2023-03-31	7.342825
2023-04-30	6.623663
2023-05-31	6.253500
2023-06-30	5.470027
2023-07-31	3.135716
2023-08-31	5.299074
2023-09-30	5.669505
2023-10-31	3.449764

2023-11-30	3.102862
2023-12-31	6.855273
2024-01-31	6.462341
2024-02-29	5.420892
2024-03-31	5.619632
2024-04-30	6.125347
2024-05-31	5.129169
2024-06-30	3.931994
2024-07-31	1.244986
2024-08-31	3.921365
2024-09-30	3.758300
2024-10-31	1.182734
2024-11-30	0.818101
2024-12-31	5.139422
2025-01-31	4.730647
2025-02-28	3.380572
2025-03-31	3.692820
2025-04-30	4.787871
2025-05-31	3.490493
2025-06-30	2.094270
2025-07-31	-0.762434
2025-08-31	2.160778

Table 5.7: the SARIMA forecast

The forecasted values can be drawn into a graph which predicts the values from March 2023 to August 2025.

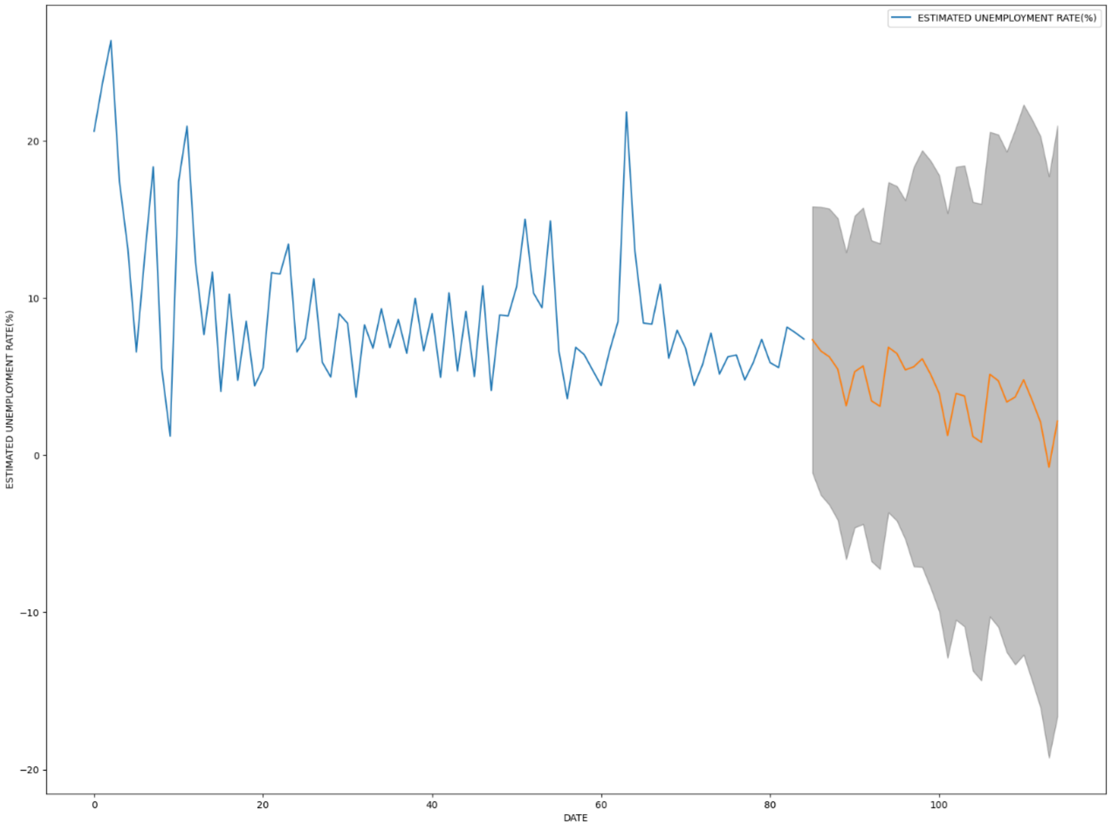

Fig 5.17: plot of the forecasted values using SARIMA model

5.1.3 Modelling and Forecasting of unemployment rate in Urban areas using Holt-Winters model

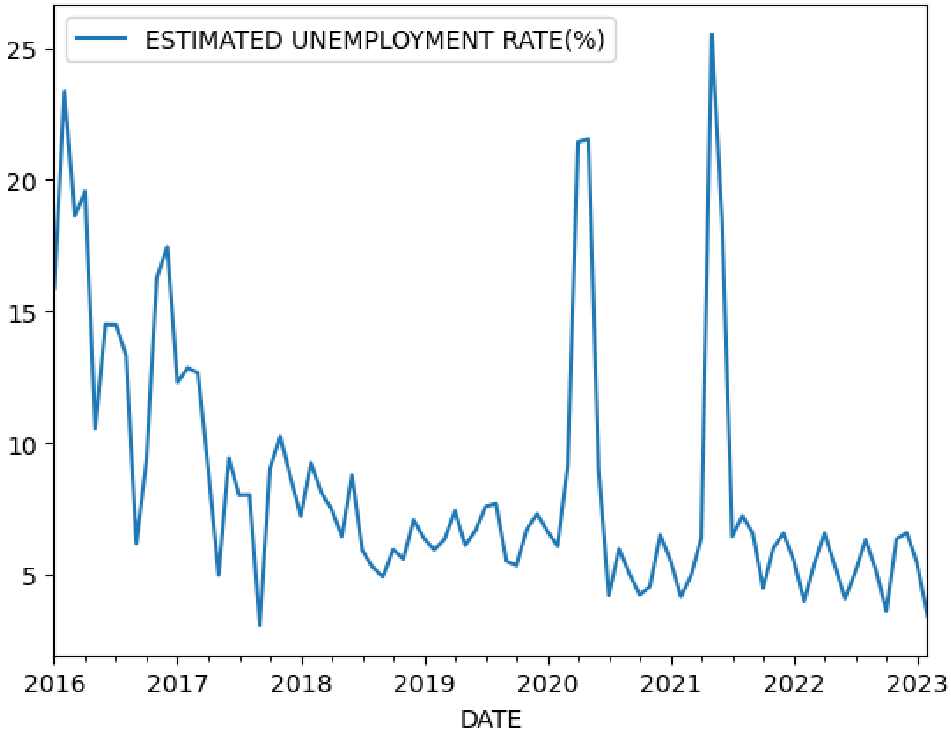

Fig 5.18: time series plot of unemployment rate from 2016 to 2023

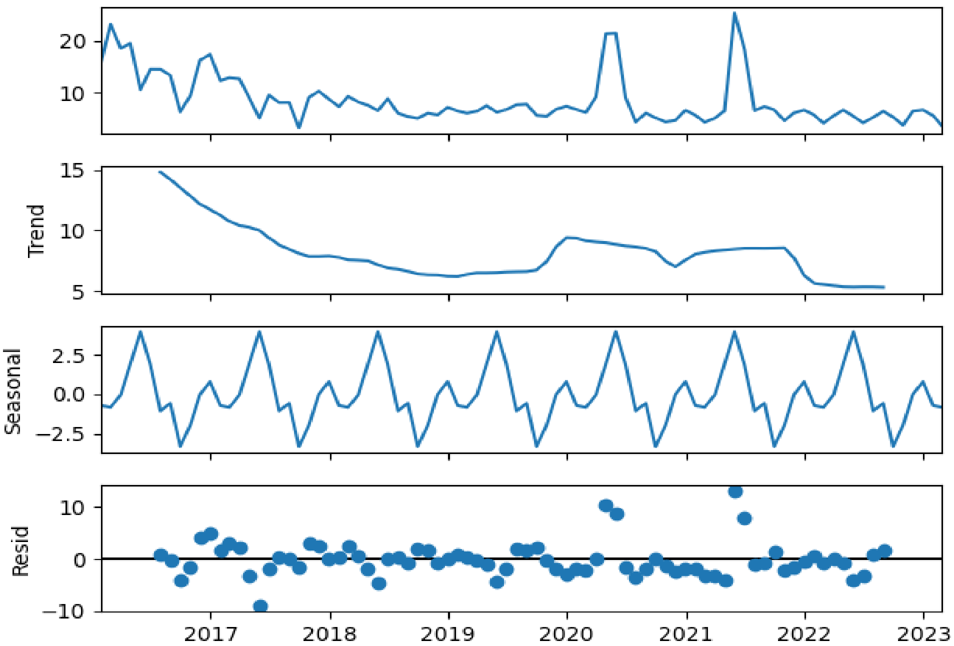

Fig 5.19: Seasonal decomposition plot

From the diagnostic plot it is clear that, the data is normally distributed. Standard residuals show how much outliers are present in the dataset. Normal Q-Q plot shows the normality of the data.

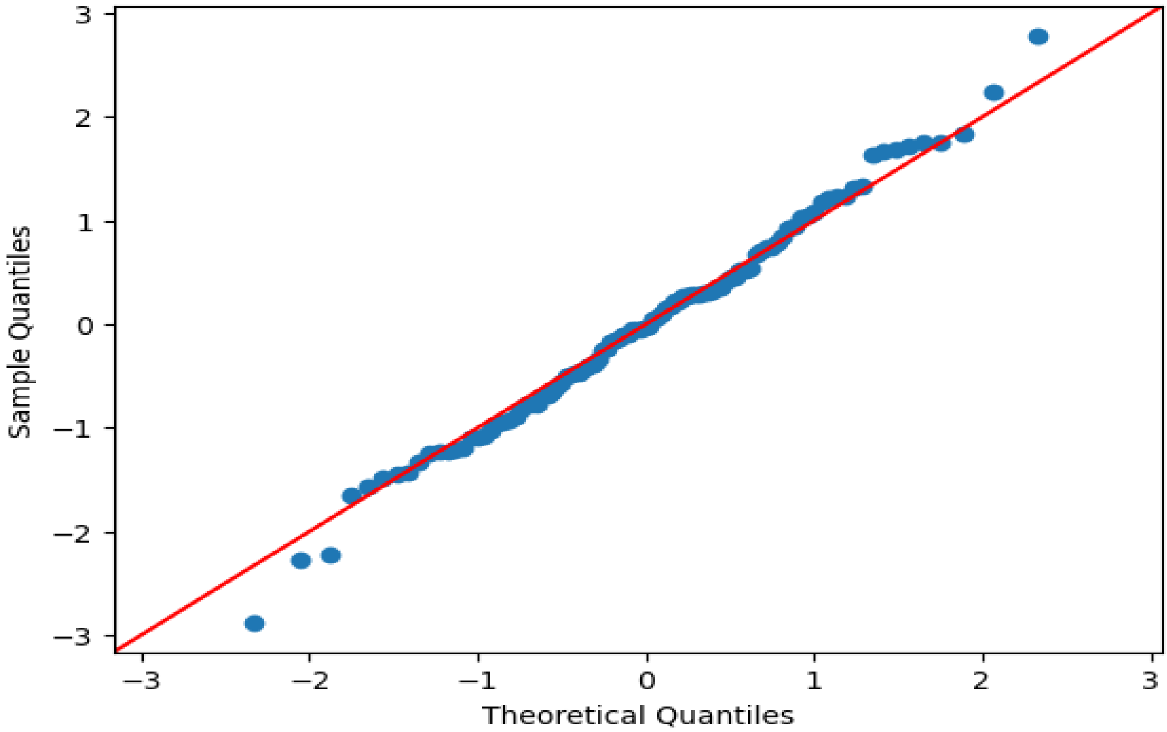

Fig 5.20 shows the Normal Q-Q plot of noise residuals of the Holt-Winters model

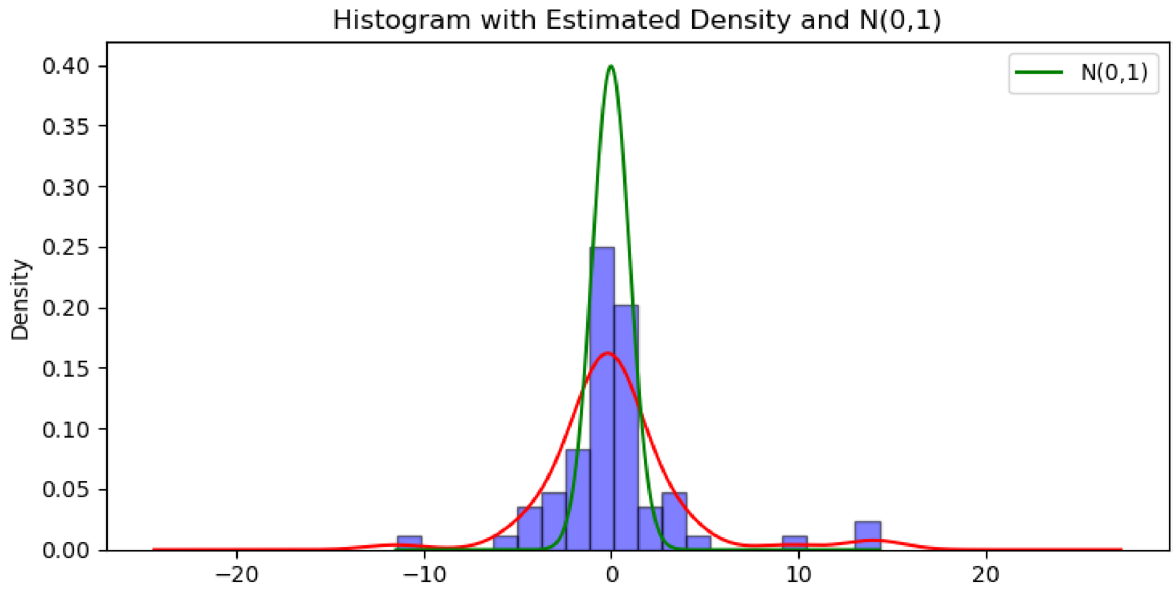

Fig 5.21: histogram plus estimated density

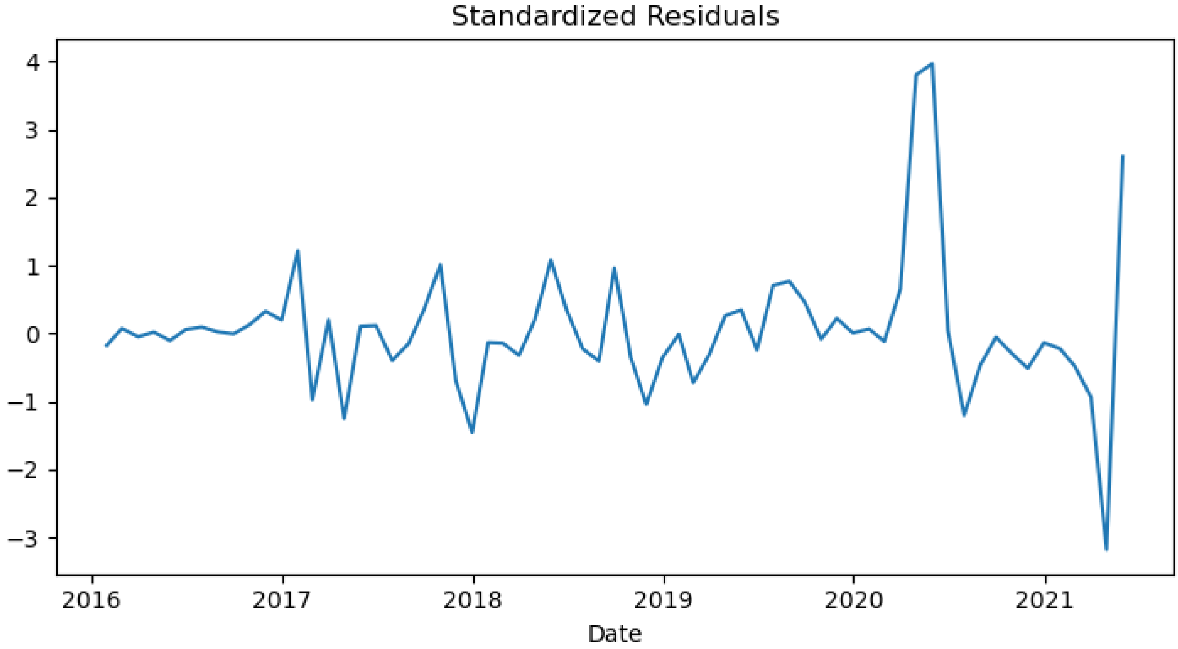

Fig 5.22: the plot of standardized residual

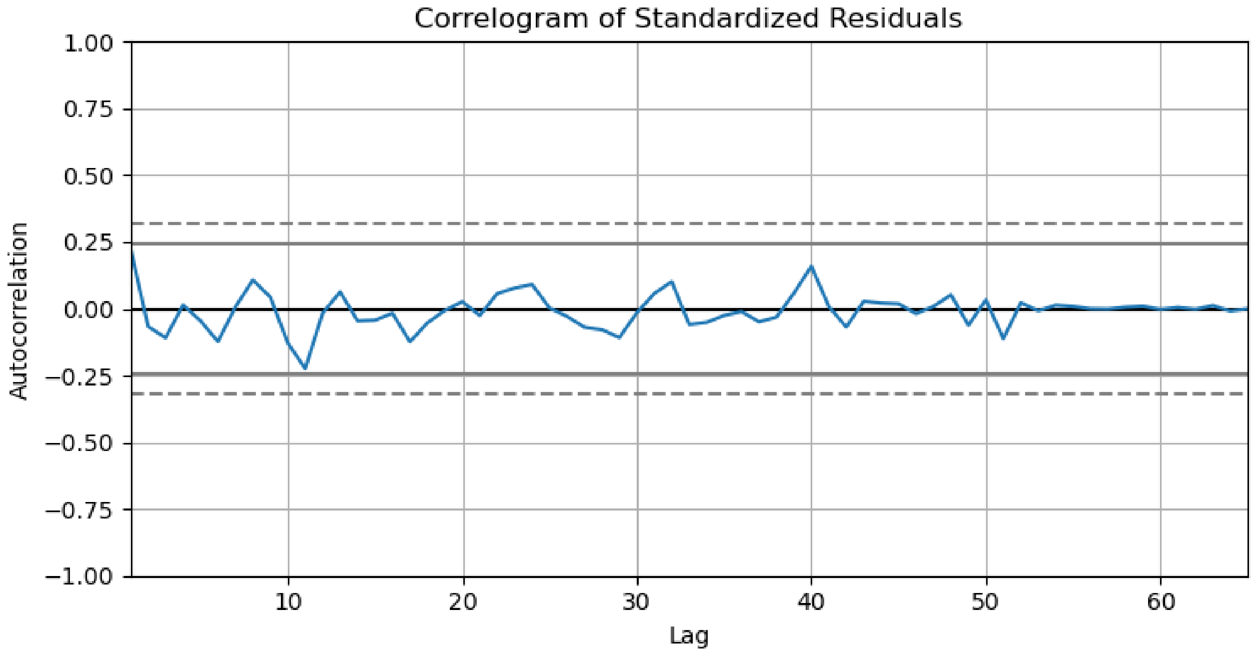

Fig 4.23: correlogram of standardized residual

Fig 5.24 shows the plot of the training data, testing data and the predicted values of the unemployment rate over time

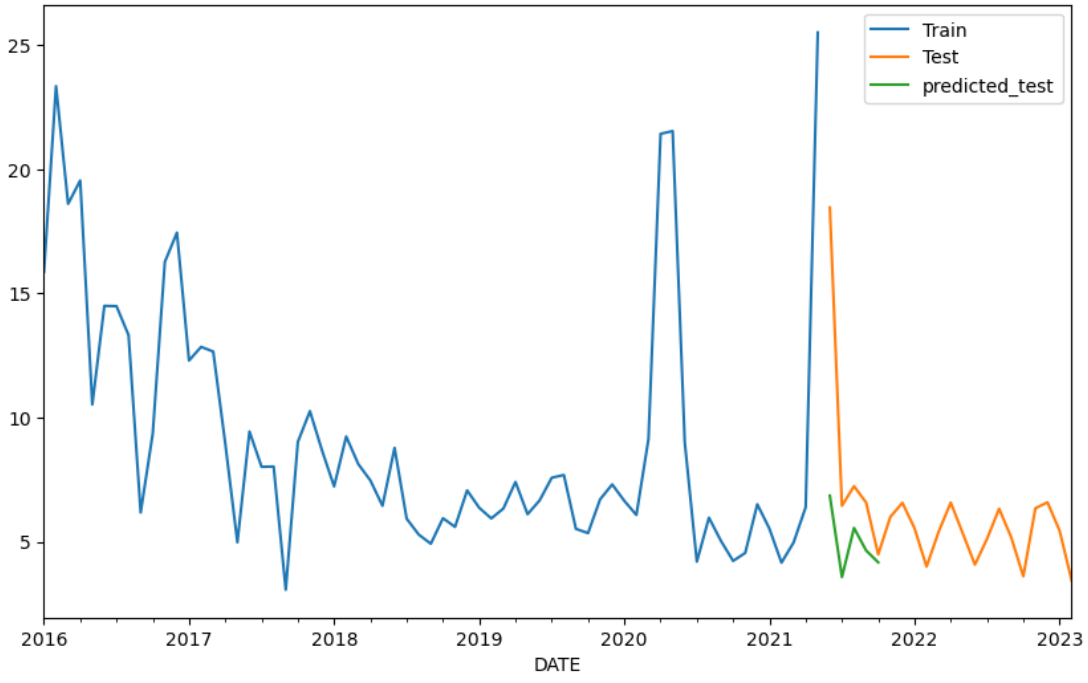

Fig 5.24

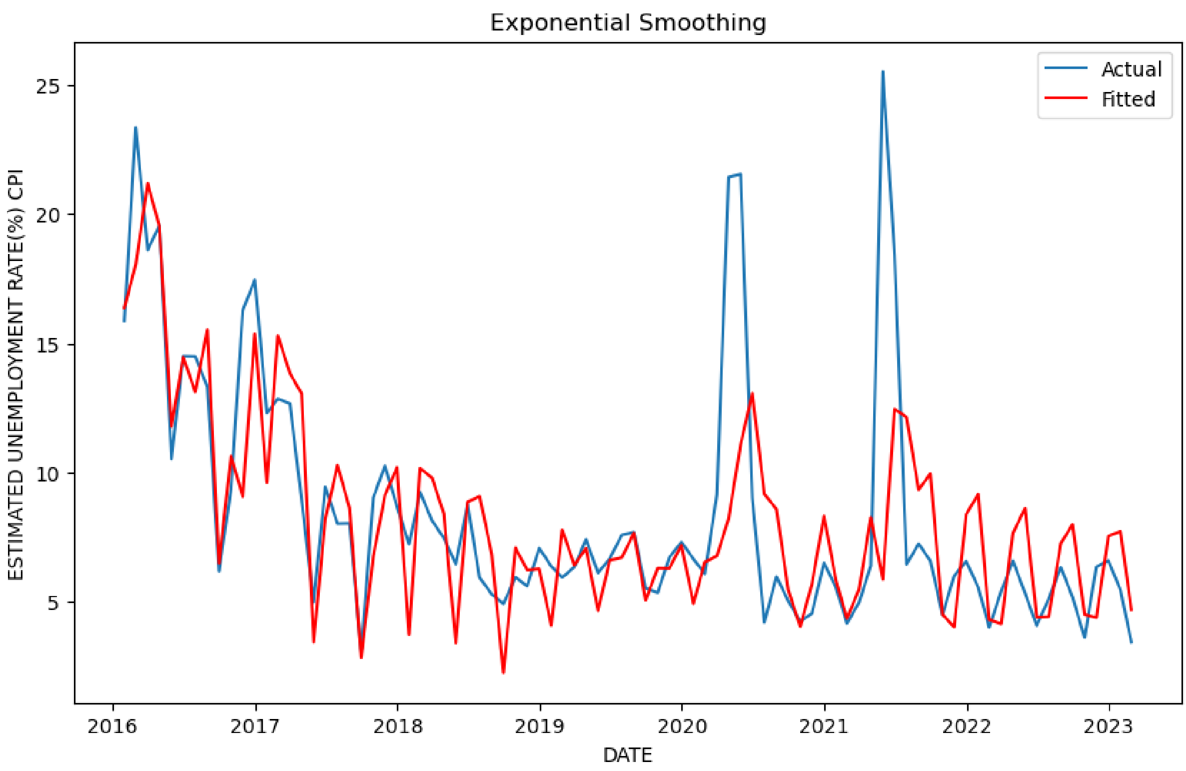

Fig 5.25 shows the fitted verses actual values plot

5.1.3.1 Forecasting of Estimated Unemployment Rate (%) using Holt-Winters model

The forecasted values along with the corresponding plot are given in table 5.8 and figure 5.20 respectively.

Date	Forecast Values	LCL	UCL
2023-03-31	4.749855	-2.08585	11.58556
2023-04-30	6.537385	-0.29832	13.37309
2023-05-31	6.695133	-0.14058	13.53084
2023-06-30	6.231506	-0.6042	13.06721
2023-07-31	4.103181	-2.73253	10.93889
2023-08-31	4.763219	-2.07249	11.59893
2023-09-30	2.194140	-4.64157	9.029849
2023-10-31	3.278819	-3.55689	10.11453
2023-11-30	4.895002	-1.94071	11.73071
2023-12-31	6.270559	-0.56515	13.10627
2024-01-31	4.444865	-2.39084	11.28057
2024-02-29	5.325001	-1.51071	12.16071
2024-03-31	5.462237	-1.37347	12.29795
2024-04-30	7.062700	0.226991	13.89841
2024-05-31	7.082503	0.246794	13.91821
2024-06-30	6.517154	-0.31855	13.35286
2024-07-31	4.313819	-2.52189	11.14953
2024-08-31	4.918544	-1.91716	11.75425
2024-09-30	2.308677	-4.52703	9.144386
2024-10-31	3.363279	-3.47243	10.19899
2024-11-30	4.957284	-1.87842	11.79299
2024-12-31	6.316486	-0.51922	13.15219
2025-01-31	4.478731	-2.35698	11.31444
2025-02-28	5.349975	-1.48573	12.18568

Table 5.8: Holt-Winters forecast values

The graph obtained by plotting the forecasted values is,

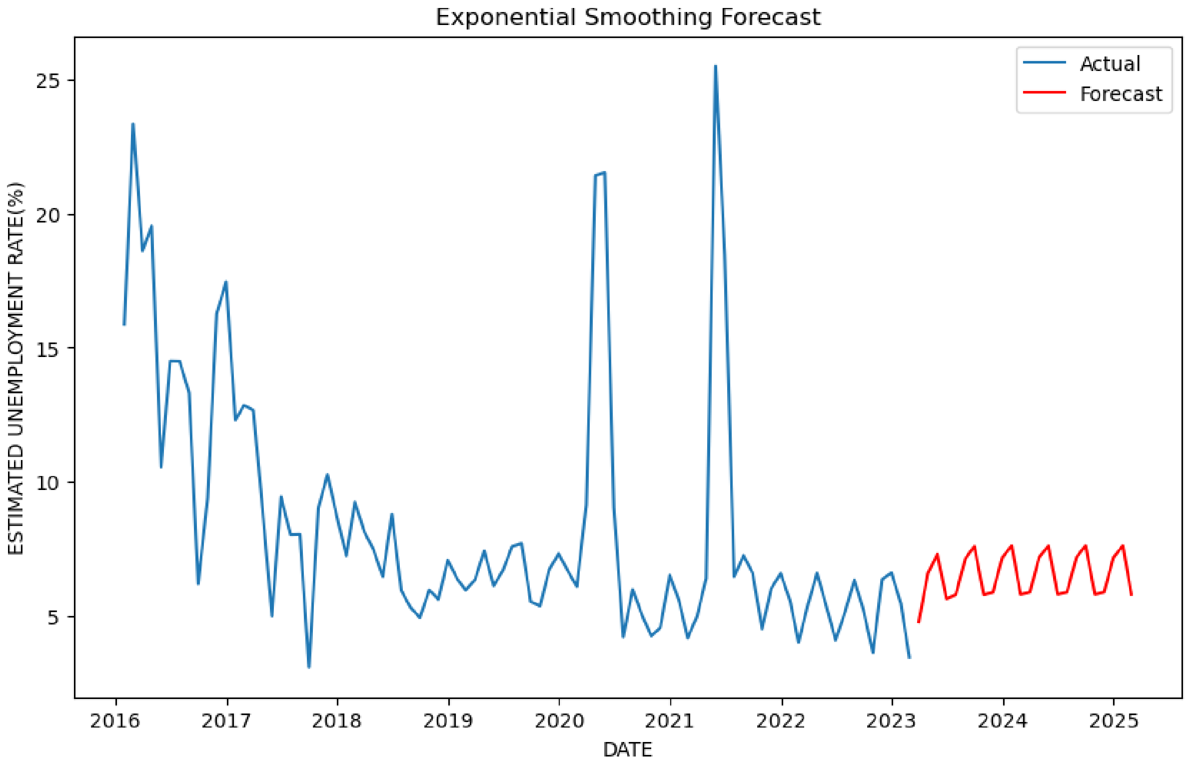

Fig 5.26: plot of the forecasted values

5.1.4 Modelling and Forecasting of unemployment rate in Rural areas using Holt-Winters model

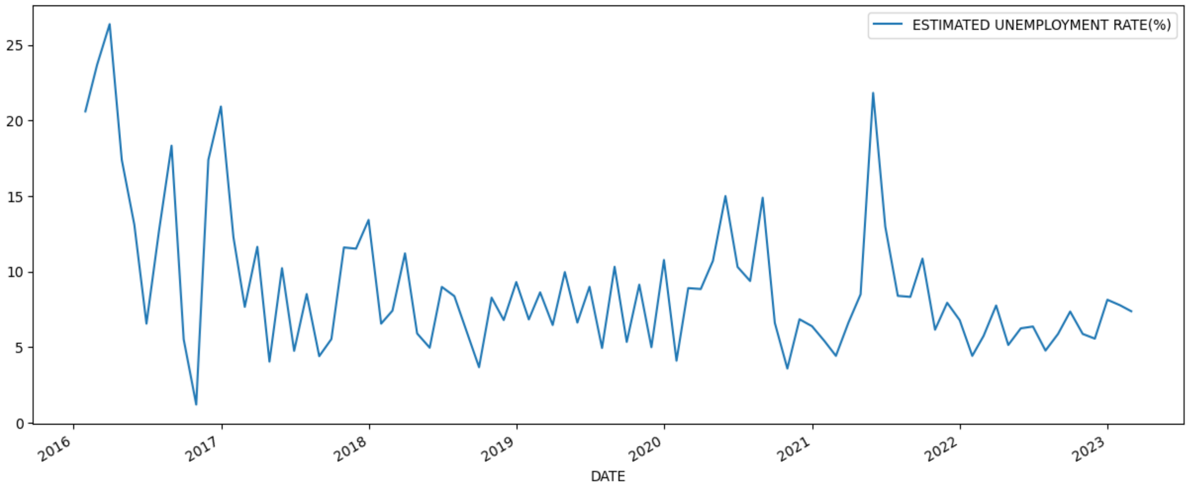

Fig 5.27 shows the time series plot of unemployment rate from 2016 to 2023

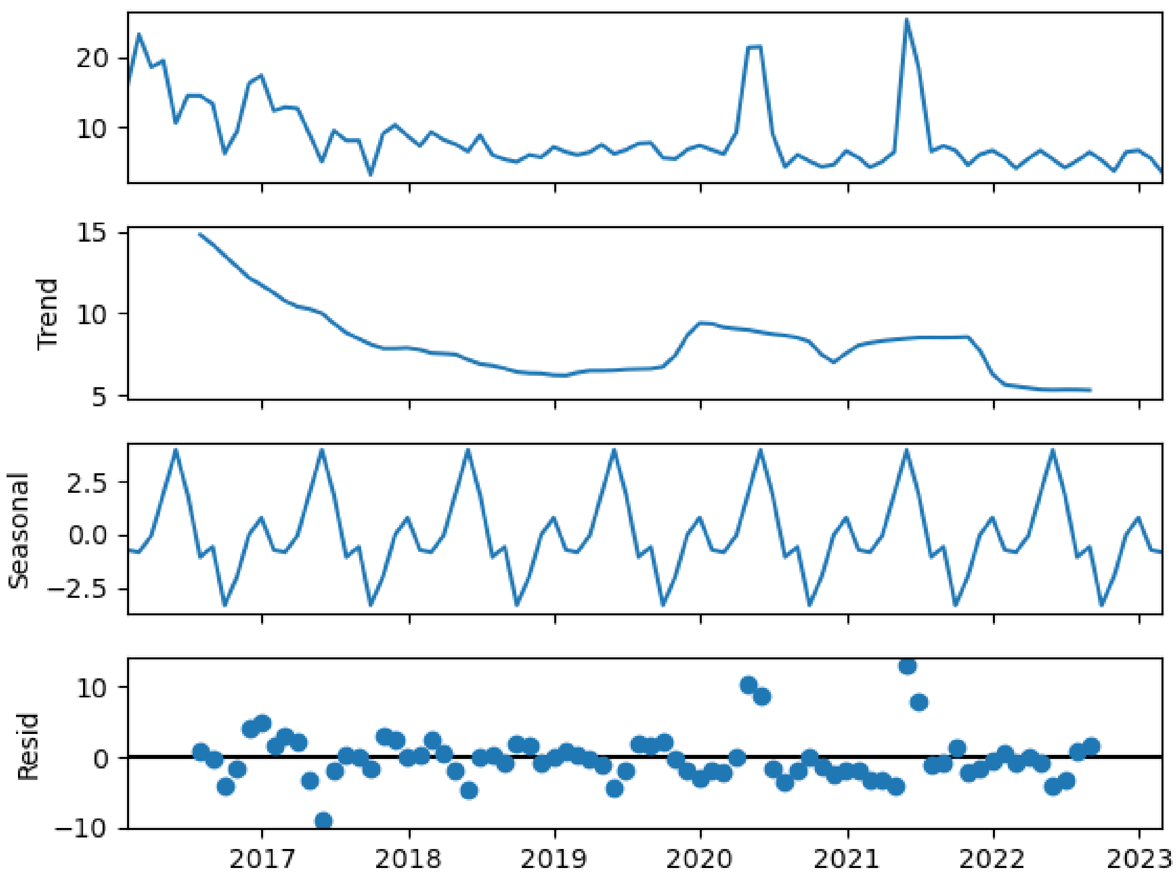

Fig 5.28: seasonal decomposition

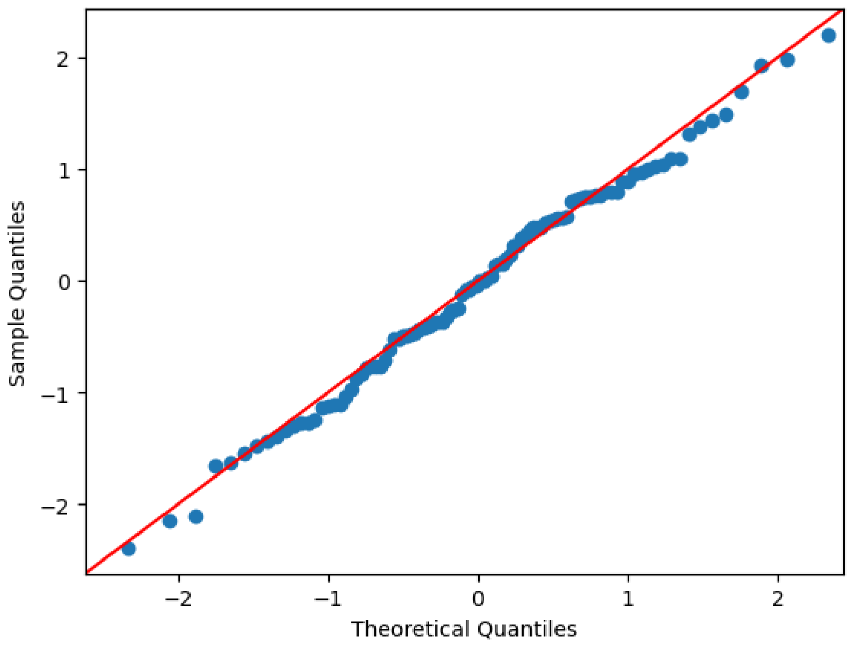

Fig 5.29: Normal Q-Q plot of noise residuals

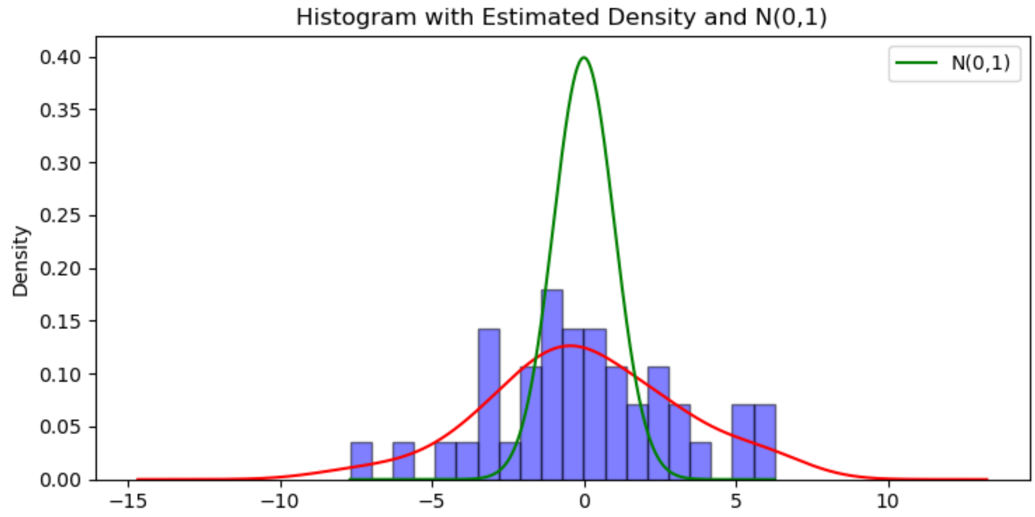

Fig 5.30: histogram with estimated density

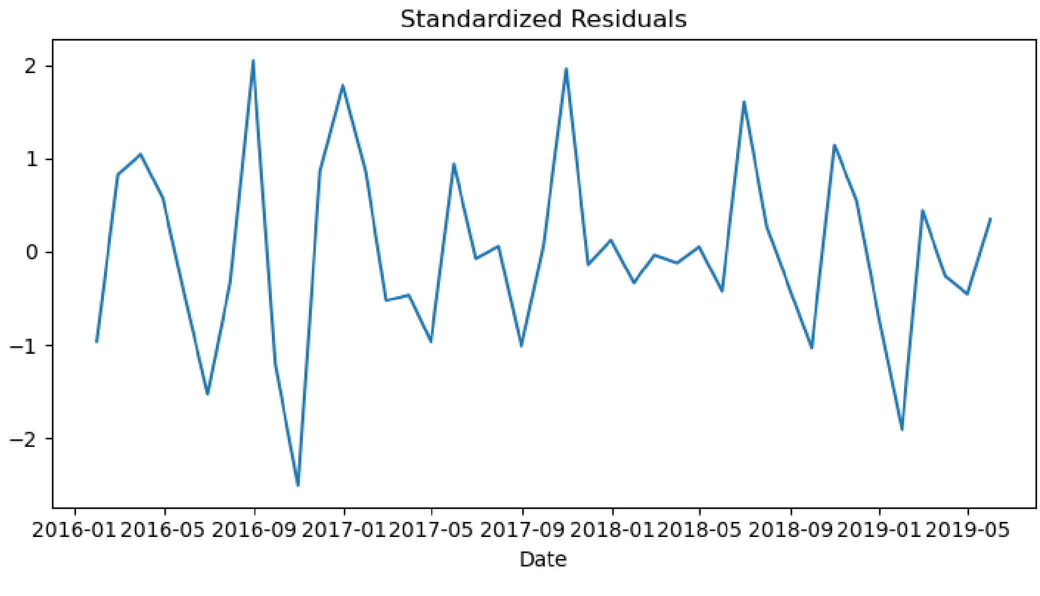

Fig 4.31: plot of standardized residual

From the diagnostic plot it is clear that, the data is normally distributed. Standard residuals show how much outliers are present in the dataset. Normal Q-Q plot shows the normality of the data.

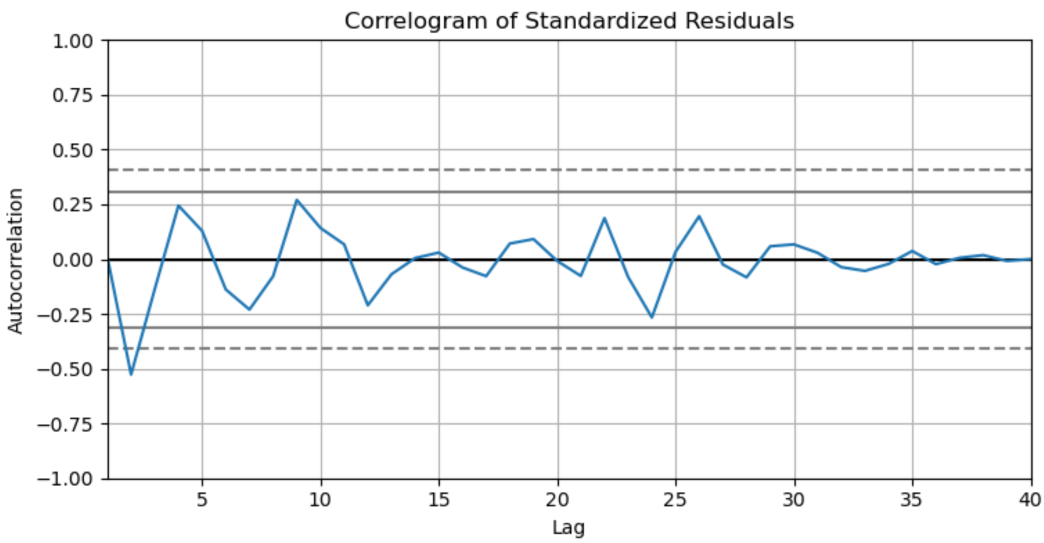

Fig 5.32: correlogram of standardized residual

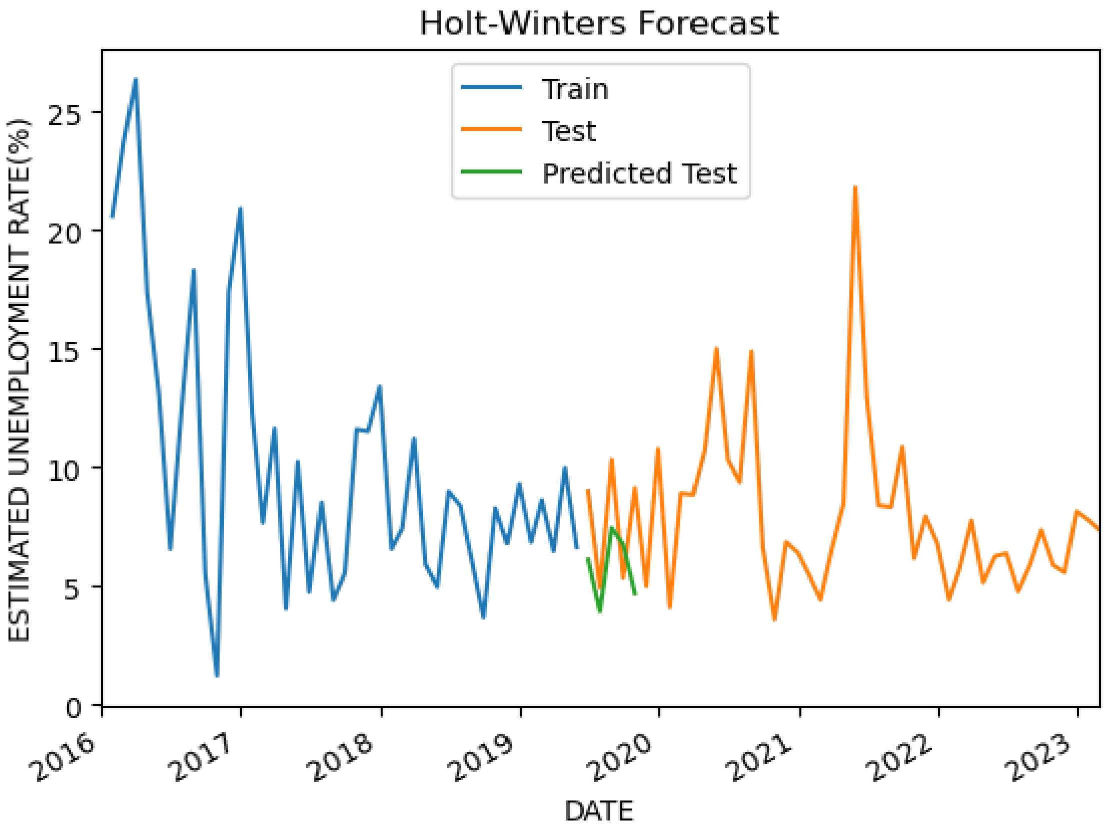

Fig 5.33: plot of the training data, testing data and the predicted values of the unemployment rate over time

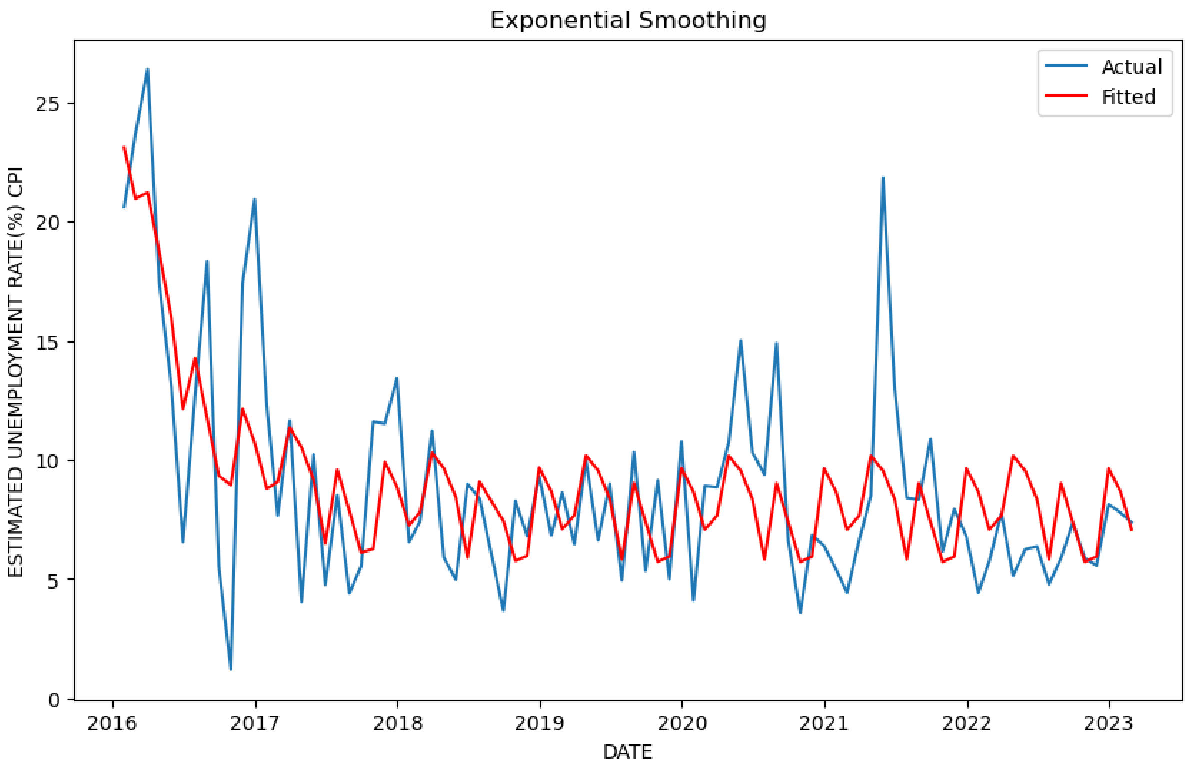

Fig 5.34: fitted verses actual values plot using the Holt-Winters model

5.1.4.1 Forecasting of Estimated Unemployment Rate (%) using Holt-Winters model

The forecasted values along with the corresponding plot are given in table 5.9 and figure 5.35 respectively.

Date	Forecast Values	LCL	UCL
2023-03-31	7.656295	2.168295	13.144295
2023-04-30	10.170427	4.682427	15.658427
2023-05-31	9.541601	4.053601	15.029601
2023-06-30	8.349093	2.861093	13.837093
2023-07-31	5.821506	0.333506	11.309506
2023-08-31	9.021621	3.533621	14.509621
2023-09-30	7.389729	1.901729	12.877729
2023-10-31	5.723613	0.235613	11.211613
2023-11-30	5.944620	0.45662	11.43262
2023-12-31	9.647503	4.159503	15.135503

2024-01-31	8.655137	3.167137	14.143137
2024-02-29	7.074795	1.586795	12.562795
2024-03-31	7.656293	2.168293	13.144293
2024-04-30	10.170425	4.682425	15.658425
2024-05-31	9.541599	4.053599	15.029599
2024-06-30	8.349092	2.861092	13.837092
2024-07-31	5.821505	0.333505	11.309505
2024-08-31	9.021620	3.53362	14.50962
2024-09-30	7.389728	1.901728	12.877728
2024-10-31	5.723613	0.235613	11.211613
2024-11-30	5.944620	0.45662	11.43262
2024-12-31	9.647502	4.159502	15.135502
2025-01-31	8.655137	3.167137	14.143137
2025-02-28	7.074795	1.586795	12.562795

Table 5.9: Holt-Winters forecast values

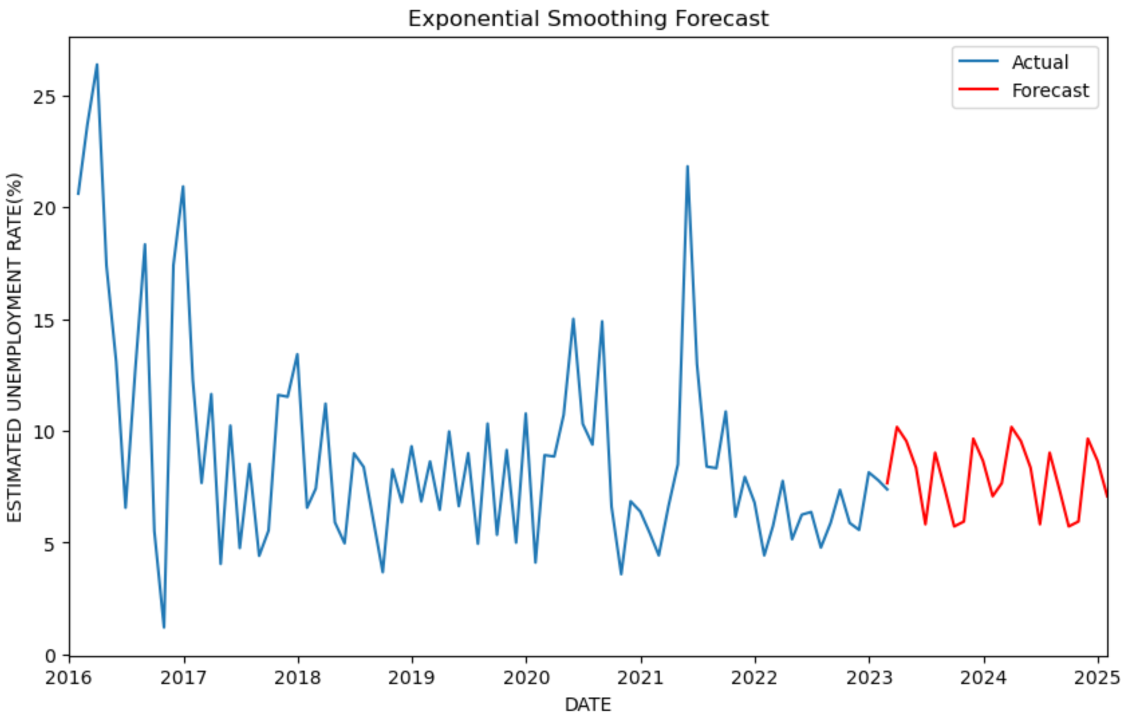

Fig 5.35: plot of the forecasted values

5.1.5 Comparison of forecast of SARIMA and Holt-Winters models

The forecast of both SARIMA and Holt-Winters models are compared with the help of MSE and RMSE values. The model having least error will be considered as the best one.

	Urban	Rural
MSE	19.87	30.76
RMSE	4.4573064989142885	5.546395307119081

Table 5.10: SARIMA model error

	Urban	Rural
MSE	12.163398421895158	7.87304100304104
RMSE	3.4876064	2.8058939757305583

Table 5.11: Holt-Winters model error

We obtain that the Mean Squared Error for Holt-Winters model is less than that of the SARIMA model in both Urban and Rural areas. Thus, the Holt-Winters model is considered as the best model.

CHAPTER 6

CONCLUSION

In accordance with the obtained results, it is evident that the error observed for the Holt-Winters model is less than that for SARIMA model. This means that Holt-Winters model can be used for the forecasting of unemployment rates in Kerala. The accuracy of the values forecasted can depend on the data quality and regular updating in the data is also necessary. From the data, it is also visible that the rate of unemployment was comparatively high in 2020 when the covid-19 pandemic hits the world.

6.1 LIMITATIONS OF THE STUDY

i) The data used in this project is secondary data.

ii) Dependence structure is not captured in Holt-Winters model.

FINDINGS AND FUTURE SCOPE

Unemployment rose to its peak during the years 2020 and 2021 when covid broke out and affected both urban and rural areas. For the best forecasting, Holt-Winters model has been chosen.

By identifying the forecasted values, one can assist the policy makers to take necessary steps in order to overcome the issue of unemployment. As the patterns get identified, it can be used for decision making. Government agencies and policymakers can make use of the unemployment rate forecasts to take necessary steps against rise in unemployment rates.

REFERENCES

1. Chakraborty, T., Chakraborty, A. K., Biswas, M., Banerjee, S., & Bhattacharya, S. (2021). Unemployment rate forecasting: A hybrid approach. *Computational Economics*, *57*, 183-201.

2. Guerard, J., Thomakos, D., & Kyriazi, F. (2020). Automatic time series modeling and forecasting: A replication case study of forecasting real GDP, the unemployment rate and the impact of leading economic indicators. *Cogent Economics & Finance*, *8*(1), 1759483.

3. Isqeel Adesegun, O., Mathew, O. O., & Omotola, S. B. (2020). Modeling Unemployment Rates in Nigeria using Time Series Approach. *Asian Journal of Mathematical Sciences*.

4. Katris, C. Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. *Comput Econ* 55, 673–706 (2020). https://doi.org/10.1007/s10614-019-09908-9

5. Khrais, I., & Al-Wadi, M. (2016). Economic growth and unemployment relationship: An empirical study for MENA countries. *International Journal of Managerial Studies and Research (IJMSR)*, *4*(12), 19-24.

6. Kramarova, K., Švábová, L., & Gabrikova, B. (2022). Impacts of the Covid-19 crisis on unemployment in Slovakia: a statistically created counterfactual approach using the time series analysis. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, *17*(2), 343-389.

7. Nkoane, S. S., & Seeletse, S. M. (2021). Forecasting unemployment rate in south africa with unexpected events using robust estimators. *International Journal of Economics and Finance Studies*, *13*(2), 199-222.

8. Sharma, S., & Soni, H. K. (2021). An unemployment prediction rate for Indian youth through time series forecasting. In *Machine Intelligence and Smart Systems: Proceedings of MISS 2020* (pp. 315-335). Springer Singapore.

9. Ullah, M., Su, K., & Jan, B. (2016). Forecasting, Cointegration and Causality Analysis of Unemployment Using Time Series Models. *Available at SSRN 2863673*.

10. Urrutia, J. D., Tampis, R. L., & Atienza, J. E. (2017, March). An analysis on the unemployment rate in the Philippines: A time series data approach. In *Journal of Physics: Conference Series* (Vol. 820, No. 1, p. 012008). IOP Publishing.