Project Report

On

PREDICTIVE ANALYSIS OF BIRTH RATES IN INDIA

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

SIFNA SUNEER

(Reg No. SM22AS020)

(2022-2024)

Under the Supervision of

MARIA NEETHU TITUS

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI – 682011

APRIL 2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **PREDICTIVE ANALYSIS OF BIRTH RATES IN INDIA** is a bonafide record of the work done by Ms. **SIFNA SUNEER** 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

Maria Neethu Titus
Assistant Professor,
Department of BCA
St. Teresa's College (Autonomous)
Ernakulam.

Nisha Oommen
Assistant Professor & HOD
Department of Mathematics and Statistics
St. Teresa's College (Autonomous)
Ernakulam.

.

External Examiners

1. ….…… 2. ……......

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **MARIA NEETHU TITUS**, Assistant Professor, Department of BCA, St. Teresa's College (Autonomous), Ernakulam and has not been included in any other project submitted previously for theaward of any degree.

Ernakulam SIFNA SUNEER

Date: SM22AS020

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 Mria Neethu Titus for the immense help duringthe 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 SIFNA SUNEER

Date: SM22AS020

ABSTRACT

An important demographic indicator that affects many facets of society, such as healthcare, education, and economic growth, is birth rates. The goal of this project is to predict India's future birth rates by combining machine learning and time series analysis methods. Policymakers can handle issues with population growth, healthcare systems, and social welfare programs by knowing and forecasting trends in birth rates.

The study starts with a thorough examination of historical data on India's birth rate, looking at patterns, seasonality, and trends. To understand the temporal dynamics of birth rates, time series analytic techniques like decomposition, autocorrelation analysis, and trend forecasting models are used. Then, using historical data and pertinent socioeconomic variables, machine learning algorithms—such as regression models, support vector machines, and neural networks—are applied to create predictive models. To guarantee accuracy and dependability in projecting future birth rates, these models undergo extensive training and evaluation procedures.

The research's conclusions deepen our understanding of India's demographic patterns and offer insightful information to scholars, politicians, and medical professionals. This project is to support proactive planning and evidence-based decision-making to meet the opportunities and problems related to population dynamics in India by utilizing cutting-edge analytical tools.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER – 1
INTRODUCTION

The study titled Birth rate trend analysis & future projections using Time series forecasting was an effort made to find out the trend of birth rates in India from the period (1971-2020) by depending on the data used in this study collected from the official website of OGD(Open Government Data) platform in India. Such trends reveal the future of our nation's birth rates which enable us to take effective measures to regulate & control as well as take preparative actions in terms of the nation's future interest.

Periodically, the calculation of the total number of live births per 1000 people in a population within a constrained or given time frame can be defined as the birth rate. The application of birth rate is essential in terms of understanding population growth, fertility trends, and the age distribution in society as the birth rate can be considered as the key demographic indicator.

High birth rates often indicate a younger population with more people entering childbearing age, while low birth rates suggest an aging population and potential issues with population growth and workforce sustainability. Factors influencing birth rates include economic conditions, cultural norms, access to contraception and family planning services, government policies (such as parental leave policies), education levels, and urbanization. The following factors such as economic conditions, cultural norms, access to contraception and national family planning services, educational aspects & urbanization as well the government policies such as Maternity leave, family planning, etc could be influenced by birth rates.

Low birth rates and high birth rates indicate population trends such that lower rates reflect an aging population and the issues that are potentially enough to affect the population growth and the sustainability of the workforce. In the case of higher birth rates, it often indicates the trend of the young population entering into childbearing age and also contradicts the aspects that lower birth rates indicate. The following factors such as economic conditions, cultural norms, access to contraception and national family planning services, educational aspects & urbanization as well the government policies such as Maternity leave, family planning, etc could be influenced by birth rates.

While we studying and analysing the pattern of countries with higher birth rates shows a general trend that the regions of such countries are at lower levels of economic development and also show lower literacy rates. Whereas in the case of countries having lower birth rates have contradicting results such as higher economic development, better literacy rates, and access to effective healthcare systems.

Historically India has experienced high birth rates due to cultural preferences for larger families, limited access to contraception, and socio-economic factors. Although there has been a gradual decline in birth rates, reflecting shifts in societal norms, improvements in health care and economic development. Various factors such as socio-economic conditions, cultural norms, access to healthcare, education and government policies has influenced the Birth rates in India.

Understanding a range of topics, including cultural norms, government regulations, socioeconomic trends, healthcarefacilities, and demographic trends, is necessary to analyze birth rates in India.
A variety of interrelated elements must be taken into consideration while assessing birth rates in India. Additionally, any analysis must take into account the intricate socioeconomic and cultural dynamics that shape the nation's population trends.

1.1 OBJECTIVES

1. To conduct Exploratory data analysis to uncover patterns, trends and insights in the underlying data structure for making informed decisions.
2. To model and forecast birth rates in India using Seasonal ARIMA model.
3. To model and forecast birth rates in India using Recurrent neural network (RNN).
4. To compare forecast by Seasonal ARIMA model and Recurrent neural network (RNN).

CHAPTER – 2
LITERATURE REVIEW

The purpose of this chapter is to give a thorough summary of all the information, theories, and research that have been done on the subject of the project. It provides the study's background, points out knowledge gaps, and argues that the effort is necessary.

Alqasemiet al. (2021) This paper presentsthe testing of two forecasting models using hospital birth data from Yemen: ANN and TS. The study makes use of four years' worth of monthly birthrate data, which is processed by each model to predict the birthrates for the next four years. The forecasting results' correctness is proven by an outcomes review.Through the explanation of the use of ANN and TS for birthrate forecasting, this study advances our knowledge of how sophisticated data mining methods can be applied to yield insightful information on population dynamics. The study's conclusions have consequences for Yemeni and other developing country decision-makers, assisting them in deciding on social services, infrastructure, and healthcare policies.

Joseph and Francis (2019) This paper provide a two-year forecast estimates for the ten regions of Ghana. In the analysis the ARIMA approach was used. Augmented Dicker-Fuller (ADF), KPSS and the Philips-Perron (PP) unit root tests were employed to test for stationarity of the series plot. KPSS (which is known to give more robust results) and PP test consistently showed that the series was stationary ($p< 0.05$) for all ten (10) regions, although there were some conflicting results with the ADF test for some regions. Tentative models were formulated for each region and the model with the lowest AIC was selected as the “Best” model fit for respective regions of Ghana.The results showed that regional variations of CS exist in Ghana. The study recommended for future studies to apply methods that will allow for forecasting for regions which failed the test under the methods used in this study.

Cantor and Land (1983) This study examines the monthly birth and death rates in the United States using seasonal ARIMA time series analysis techniques. For model selection, the conventional Box-Jenkins diagnostic checks are used.As a result, the final models for both series have seasonal moving-average components and strong second-order autoregressive components. Furthermore, the birth rate data provides some support for weekly periodicities.

Leung (1995) The purpose of this study is to forecasting births, with the main focus on univariate time series methods. The demographic renewal equation for births serves as the foundation for the development of a general autoregressive integrated moving average model for birth time series. A thorough examination is conducted of the four-stage Box-Jenkins modelling process, which includes model identification, estimation, diagnostic, and forecasting. Australian birth time series are modelled and forecasted using this methodology. Lastly, a comparison is conducted between cohort component projections of births for Australia and time series forecasts.

Ali and Khan (2019) The study aims to forecast Pakistan's population using a variety of demographic factors that have a substantial impact on population size. Using deterministic techniques and ARIMA (Auto Regressive

Integrated Moving Average) models, the data was examined for several Pakistani demographic indicators. The two approaches have been compared in order to evaluate how accurate their population estimates are. Using the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Akaike Information Criterion (AIC), it was discovered that ARIMA outperforms the Simple Exponential Smoothing (SES) method in population forecasting for Pakistan. As a result, it was determined that the ARIMA model is helpful and that using it to forecast Pakistan's population is advised.

Djagbletey et al. (2019) The purpose of this study was to evaluate the seasonal SARIMA model against the Holt-Winters seasonal forecasting techniques using an 11-year time series of birth data.Through the use of various mathematical models, investigations have demonstrated periodic fluctuations in the frequency of births.It has shown periodic variations in the number of births using different mathematical models.Seasonal Autoregressive Integrated Moving Average (SARIMA) model was obtained on a monthly number of births for an 11-year data to ascertain which forecasting technique would best explain the data, this study did not, however, compare the developed model with alternative approaches. The purpose of this study was to evaluate the seasonal SARIMA model against the Holt-Winters seasonal forecasting techniques using an 11-year time series of birth data.Using the birth data, Holt-Winters and seasonal ARIMA forecasting techniques were used. Out of all the models, SARIMA (2, 1, 1) x (1, 01,) had the lowest in-sample forecasting errors. When compared to the Holt-Winters form of additive and multiplicative approaches based on the forecasting accuracy indices of the monthly number of births for an 11-year period, the out-of-sample errors also demonstrated that all of the SARIMA models had fewer errors. The Holt-Winters models were inferior to the SARIMA models.

Belaghi et al. (2021) The study aims to predict preterm birth in nulliparous women using logistic regression and machine learning. In order to predict overall and spontaneous preterm birth, they developed logistic regression and machine learning models in a "training" sample using data from the first and second trimesters. In an independent "validation" sample, we evaluated the model's performance using a variety of accuracy metrics, such as sensitivity, specificity, positive predictive value, negative predictive value, and area under the receiver operating characteristic curve (AUC).Despite offering somewhat better AUC than logistic regression, artificial neural networks were still unable to accurately predict preterm birth in the first trimester. However, both machine learning and logistic regression techniques showed a moderate improvement in prediction when second trimester data was included.

Masha (2020) In this study, he looked at Ghana's birth rate on a monthly basis using the Box-Jenkins approach.The Ghana Statistical Service provided him with monthly birth rate information for the study period spanning from January 2014 to December 2019.He found that the ARIMA (1, 1, 1) model fit the data the best out of all the models investigated in the research, with the lowest normalized Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) values. The 12-month birth rate for the year 2020 was subjected to the produced model. The ARIMA (1, 1, 1) model was shown to have the best fit among the models examined in the study, with the lowest normalised Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. Ultimately, a negative correlation was found.

Zakria (2009) The goal of the study is to use a stochastic ARIMA model to anticipate Pakistan's population trend over the next 20 years based on historical trends.This study models the population of Pakistan from 1951 to 2007 using the Box Jenkins ARIMA technique.He gathered population statistics for Pakistan from 1951 to 2007 for this purpose.Moreover, a 20-year population projection for Pakistan is made using the frugal ARIMA (1, 2, 0) model. If the current trend of growth rate is maintained, he calculated that Pakistan's population will reach roughly 230.7 million by 2027.The population predicted by the parsimonious ARIMA model is in close agreement with the population forecasted by several bureaus in Pakistan. 229 million people were believed to live in these bureaus.

Katerina (2024) The study aims to improve the modelling and forecasting results by combining a linear ARIMA method with an intelligent non-linear method, the artificial neural network ANN. There are three stages to time series analysis: characterization, modelling, and forecasting. ARIMA, ANN, and the suggested ARIMA-ANN approach are used to conduct the analysis. Unstructured data sources are examined using data mining, sentiment analysis, and natural language processing (NLP) approaches. Both the intelligent and quantitative models produce acceptable outcomes, but the hybrid model performs significantly better across the board. The hybrid approach produced results with a high degree of efficiency based on the measurement indicators of the prediction performance. Compared to the ARIMA and ANN models, the ARIMA-ANN model yielded substantially better outcomes. These outcomes come from combining the two methods, which leads to an accurate identification of the various data mode structures. By creating a new forecast model called hybridization, this study advances the fields of time series forecasting in Albania and intelligent or quantitative modelling.

CHAPTER - 3
MATERIALS AND METHODOLOGY

3.1 DATA DESCRIPTION

The dataset contains the crude birth rates in India from the year 1971 to 2020, categorized by total, urban and rural areas on yearly basis. Data used in this study is collected from Open Government Data Platform and Sample Registration System.

3.2 METHODOLOGY

The initial and vital step in this study analysis was study the data in detail. The main purpose of the study was to forecast the future birth rates in India using time series model and machine learning model. First, the model was forecasted using time series technique, then the model was forecasted using machine learning technique. Finaly compare the forecasting by the two methods by calculating MSE and RMSE values of both models.

➢ TOOLS FOR ANALYSIS & FORECASTING

- Exploratory data analysis
- Seasonal ARIMA Model
- Recurrent Neural Network (RNN)

➢ TOOL FOR COMPARISON

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

CHAPTER – 4

EXPLORATORY DATA ANALYSIS

EDA is an important step in the data analysis process that involves searching and understanding datasets before using statistical methods to estimate the model. The goal of EDA is to identify the structure and characteristics of the data before formal modelling or hypothesis testing.

Next, Using EDA to check for null and missing values, summarize and visualize the data to gain a deeper understanding of it. Compiling summary statistics like mean, median, mode, standard deviation, and soon may fall under this category. This also includes displaying the data and breaking it down in to trends, seasonal patterns, and residual data.

1. Visualizing time series data using line plots to understand its distribution, trends and variability over time.
2. Using summary statistics to characterize the data's central tendency and distribution , such as mean, median, standard deviation and percentiles.
3. To find seasonality and dependence patterns in the data, analyze the autocorrelation and partial autocorrelation functions.
4. Evaluating stationarity using statistical tests such as Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test or by visually examining rolling statistics charts.

EDA is the first and most important step in data analysis to understand the data and find the missing or null values. It also give an insight about which model are need for the dataset to forecast accurate results.

CHAPTER – 5
TIME SERIES

5.1 TIME SERIES ANALYSIS

Time series is the sequence of data points that is collected over a period of time and ordered chronologically. The key characteristic of time series is that it is indexed in time order. If we want to plot time series data on a graph, one of the axis would always be time. These data points are past values which are analyzed to forecast future trend. If a random variable X is indexed to time t, the observation $\{X_t, t \in T\}$ is called a time series, where T is the index set. Time is one of the component of this data hence it is time dependent.

The main components of time series are:

- **Trend :**it is a long term movement of a time series. It could be increasing, decreasing or stable.
- **Seasonality :**It is the repetitive and predictable patterns that appear in the data at regular periods with period of cycle within one year.
- **Cyclic variation :**cyclic variations are gradual, relatively long term, up and-down repetitive movements of the time series.fluctuations in the data that occur over extended periods but do not have fixed periods like seasonality.
- **Random variation:**This represents the random or erratic fluctuations that can't be explained by the trend, seasonality, or cyclic components. These irregular variations can result from unforeseen events, measurement errors, or other unpredictable factors.

A given time series can be

5.2 STATIONARY TIME SERIES

A stationary time series is characterized by constant mean, variance, and autocorrelation structure over the course of the entire series. This suggests that the fundamental mechanisms producing the data stay constant across time, which facilitatesto approximate and forecast future values. The application of methods like autoregressive integrated moving average (ARIMA) models, which presuppose a stable environment, is frequently made possible by stationary time series.

5.3 NON-STATIONARY TIME SERIES

Non-stationary time series exhibit fluctuating means, trends, and other statistical characteristics that change with time. Seasonality, outside influences, or intrinsic instability in the data-generating process are some of the possible causes of these variations. Modelling and prediction can be difficult with non-stationary data because developing patterns may not be well captured by conventional techniques. Differentiating the data into a stationary form is a typical method of addressing non-stationarity. To eliminate trends or seasonality, differencing entails deducting each data point from its lagged counterpart. Once converted, the data can frequently be efficiently modelled utilizing stationary approaches.

5.4 AUTOCORRELATION FUNCTION

The Autocorrelation Function (ACF) is a graphical tool used to evaluate the correlation between a time series and its lag variants. Stated differently, it measures how similar a time series values are at various points in time. Plotting the ACF allows one todetermine if the present observation and observations made at different lags have a substantial association. There may be a seasonality component associated with a certain lag if there is a significant decline in autocorrelation following that lag. A declining ACF over time suggests that there may be an AR component in the time series. In a model such as ARIMA, the ACF plays a crucial role in determining the order of the moving average (MA) term.

5.5 PARTIAL AUTOCORRELATION FUNCTION

The Partial Autocorrelation Function (PACF) extends the concept of autocorrelation by capturing the direct correlation between two observations while accounting for the indirect correlations introduced by the intermediate lags. It essentially measures the correlation between the current observation and observations at specific lags, after removing the effects of the intervening lags. By plotting the PACF, analysts can identify the lag values where the direct correlation is significant. A sharp cutoff in the PACF after a certain lag indicates a potential autoregressive (AR) component in the time series. Just like the ACF, the PACF assists in determining the appropriate parameters for models like ARIMA.

5.6 ARIMA (AUTOREGRESSIVE INTEGRATED MOVING AVERAGE)

ARIMA which stands for Aoto – Regressive Integrated Moving Average, is a powerful time series forecasting model used to capture patterns, trends, and dependencies present in time series data. ARIMA combines three main components: autoregression (AR), differencing (I for integrated), and moving average (MA) .

Auto-Regressive (AR) component : The AR component represents the correlation (autocorrelation) between an observation and a specific number of lag observations.It illustrates the series' linear dependence between its previous values and current value. represented by the parameter p, which indicates how many lag-time observations are incorporated into the model.

Differencing (I) Component : The series' differencing to make it stationary is represented by the I component. Because many models assume that the statistical features of the series do not vary over time, stationarity is crucial in time series analysis. It is represented by the parameter d, which indicates the quantity of differences required for the series to become stationary.

Moving Average (MA) component : The MA component models the relationship between the current observation and a specified number of past errors or residuals. It helps capture the impact of past shocks on

the current value. The order of theMA component is denoted as "q." The ARIMA model is often denoted as ARIMA (p, d, q), where "p," "d," and "q" are the orders of the AR, differencing, and MA components, respectively.

5.7 SARIMA (SEASONAL ARIMA)

SARIMA, or Seasonal ARIMA, extends the basic ARIMA model to handle time series data with both non-seasonal and seasonal patterns. This is especially useful for data that exhibit recurring patterns at fixed intervals, such as sales data affected by yearlyseasonality. SARIMA includes additional components to capture the seasonal variations:

Seasonal Auto-Regressive (SAR) Component: This component models the relationship between the current observation and past observations at the same lag in previous seasons. It captures the effect of seasonality on the data. The order of the seasonal AR component is denoted as "P."

Seasonal Differencing (S) Component: Similar to the non-seasonal differencing, the seasonal differencing component is used to remove the seasonality from the data. The order of seasonal differencing is denoted as "D."

Seasonal Moving Average (SMA) Component`: This component models the relationship between the current observation and past errors or residuals at the same lag in previous seasons. It accounts for past shocks that persist across seasons. The order of the seasonal MA component is denoted as "Q." SARIMA is denoted as SARIMA (p, d, q) (P, D, Q, s), where "s" represents the season's length.

5.8 KWIATKOWSKI-PHILLIPS-SCHMIDT-SHIN TEST

The KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test is a statistical test used to determine if a time series is stationary or non-stationary. It assesses whether the variance of a time series is constant over time, with the null hypothesis being that the series is stationary.

5.9 ROLLING STATISTICS PLOT

It is a graphical representation used in time series analysis to visualize the rolling or moving average and standard deviation of a time series data.It consists of original time series data overlaid with lines representing the rolling average and rolling standard deviation.

Rolling average :A set window of consecutive data points is averaged to determine the rolling average. Every time a new set of data points is collected, the average is computed as the window "rolls" or travels along the time series.This evens out transient variations and draws attention to long-term patterns or trends in the data.
Rolling standard deviation :the rolling standard deviation is computed by taking the same fixed window of consecutive data points' standard deviation.It calculates the data's variability or dispersion within that window.It aids in detecting variations in variability over time, much like the rolling average.

5.10 AKAIKE INFORMATION CRITERION (AIC)

The AIC is a method for examining how well a model fits the data. In statistics, AIC is used to compare different models and determine which one is best fit for the data. AIC does not test hypothesis but estimates information loss when a model represents data. Its formula, AIC = 2k-2ln(L), consider the parameters (k) and likelihood function (L). Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

5.11 FORECASTING

Forecasting is the process of making predictions based on past and present data collected at regular intervals over time. Time series forecasting covers analysing historical patterns, identifying trends and seasonal fluctuations. Also use statistical models or machine learning algorithms to make predictions.

CHAPTER – 6
RECURRENT NEURAL NETWORK

Recurrent neural network (RNN) is a type of artificial neural network is made to handle sequence data by maintaining a memory of past inputs. Recurrent neural networks (RNNs) display dynamic temporal activity because of connections that create directed cycles, in contrast to standard feedforward neural networks that process input data independently. The components of RNN are,

Recurrent connections :RNNs can retain information over time because of their self-looping connections. They are able to record dependencies between sequence items as a result.

Hidden State: An RNN keeps track of a hidden state vector at each time step that contains an overview of the data it has seen thus far. The network is able to recall historical data because this hidden state is modified in response to both the current input and the previous hidden state.

Time Unfolding:RNNs can be thought of as being unrolled over time, with a layer in the unfolded network being represented by each time step. This makes it possible to train using conventional backpropagation algorithms through time (BPTT).

Processing sequences :RNNs work effectively when processing sequences with varying lengths. They are helpful for a variety of applications, including time series prediction, speech recognition, natural language processing, and more since they can analyze inputs of various lengths and generate outputs at each time step.

The vanishing gradient problem, on the other hand, affects conventional RNNs because during training, gradients exponentially decrease as they propagate back in time. This may make it more difficult for the network to identify long-term dependencies in sequential data. Many RNN variations, such as Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks, have been created to address this problem. These architectures include gating features that facilitate more efficient learning of long-range relationships and help alleviate the vanishing gradient problem.In general, RNNs and its variations have developed into strong instruments for sequential data analysis, making a variety of tasks possible, including sentiment analysis, machine translation, language modelling, and more.

Time series forecasting tasks can be successfully performed using Recurrent Neural Networks (RNNs). This is how it usually operates:

Data Preparation: A format appropriate for input into the RNN is created using time series data. In order to do this, the data are usually formatted into fixed-length sequences, each of which contains past observations that have led to the goal forecast point.

Model Architecture: The particular forecasting goal and data properties are taken into consideration when selecting the RNN architecture. Basic RNNs, Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) are popular options. Because they can alleviate the vanishing gradient problem and capture long-term dependencies, LSTM and GRU networks are frequently chosen.

Training: Using historical time series data, the RNN model is trained. Using optimization methods like stochastic gradient descent (SGD) or its derivatives, the model is trained to map input sequences to target values by modifying its internal parameters.

Validation: To determine if the model can generalize to previously untested data, its performance is assessed on a different validation set after training. The performance of the model can be measured using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), or others.

Forecasting: The RNN model can be used to produce forecasts for upcoming time steps after it has been trained and verified. In order to accomplish this, the model iteratively predicts future values one step at a time using previous data as input. Forecasts that span several time steps ahead of time can be created using the anticipated values.

Evaluation: Lastly, the RNN model's forecasts' accuracy is assessed using the relevant metrics. It is imperative to evaluate the model's performance over various forecast horizons and juxtapose it with alternative forecasting techniques or baseline models.

All things considered, RNNs provide a strong and adaptable foundation for time series forecasting applications, enabling models to effectively represent intricate temporal patterns and dependencies in the data. RNN-based forecasting is dependent on a number of variables, including model architecture, hyperparameter tuning, data quality, and the availability of enough training data, much like any other machine learning technique.

CHAPTER - 7
ANALYSIS

The analysis chapter of this study digs in to two important techniques used for forecasting and understanding patterns in data: SARIMA modelling and Recurrent Neural Network.

EDAis an important step in data analysis process that involves exploring and understanding a dataset before applying statistical techniques predictive models. It helps to identify patterns, uncovering potential correlation between variables, detect anomalies, and guiding to select statistical techniques for further analysis.

SARIMA model is used for time series forecasting that takes in to account both seasonal and non-seasonal patterns in time series data. It helps in forecasting future values based on historical data by capturing and modelling various components of time series data such as trends, seasonality and noise.

RNN is used as a powerful tool for time series forecasting and it has the ability to learn from historical data to generate accurate predictions for future time steps.

The goal to achieve through this technique is to gain insights from the data, to make predictionsaccuratelyand to discover hidden patterns.

7.1 URBAN BIRTH RATE

7.1.1 EXPLORATORY DATA ANALYSIS

Initially, for the analysis we do exploratory data analysis.

Table 7.1 shows that there is no null value in the given data

Birth rate urban	**0**
dtype	**Int64**

Table 7.1

Table 7.2 shows the data summary

	Birth rate urban
Minimum	**16.100000**
1st Quartile	**18.525000**
Mean	**22.798000**
Median	**22.150000**
S D	**4.628593**
3rd Quartile	**27.600000**
Maximum	**30.500000**

Table 7.2

The initial step of time series analysis is to draw a time series plot. The time series plot of birth rates in India is given in figure 7.1

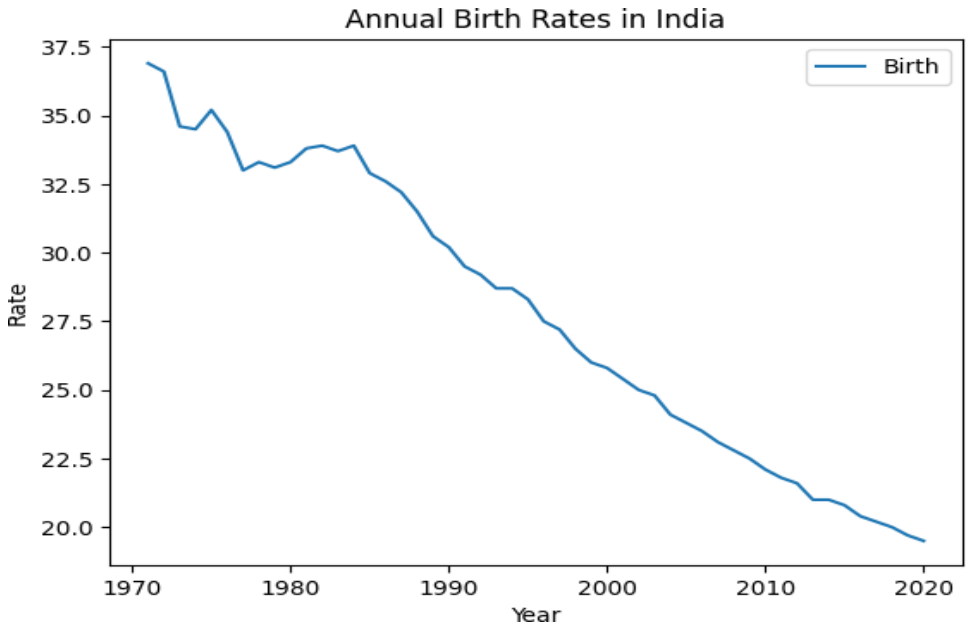


Figure 7.1

7.1.2 TIME SERIES

7.1.2.1 DECOMPOSITION OF TIME

Do seasonal decomposition for evaluating the trend, seasonal and random components of the time series and draw the seasonal plot. The seasonal plot is given in figure 7.2

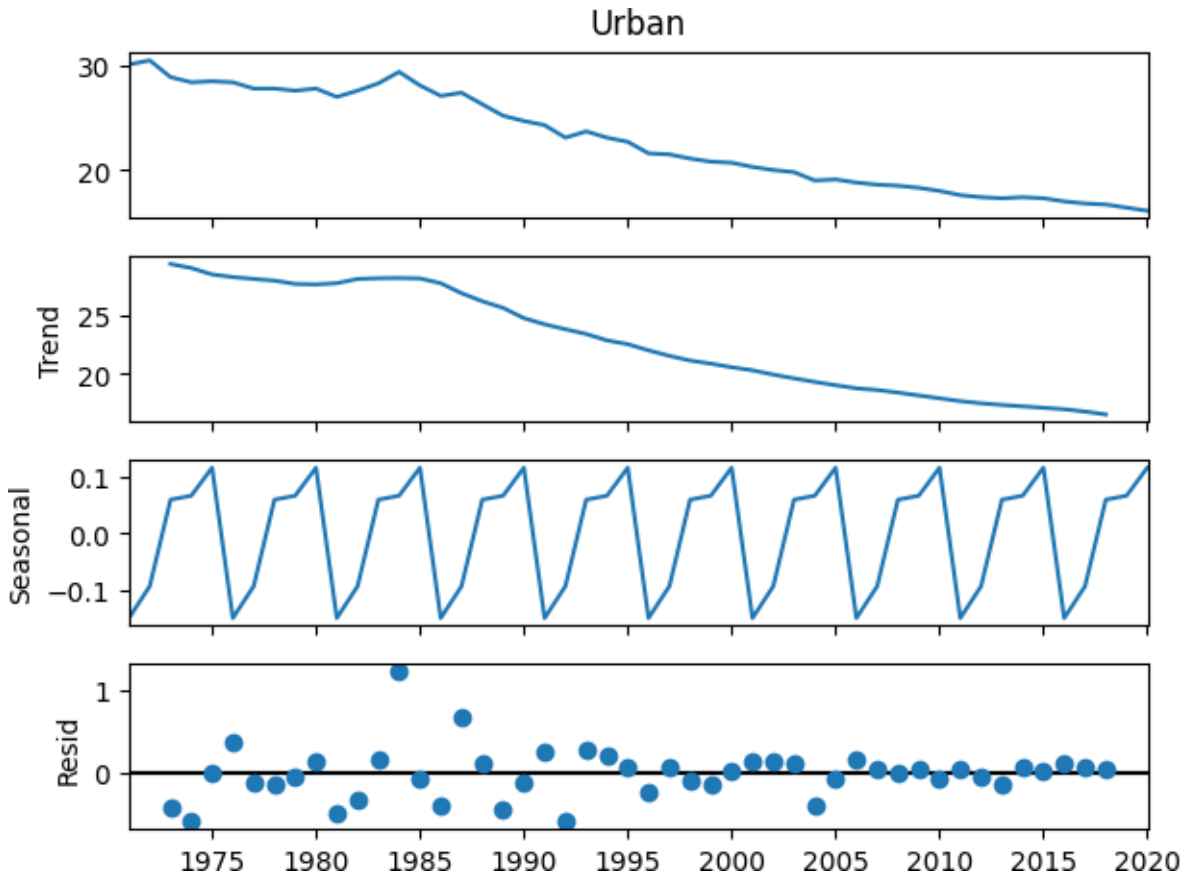


Fig 7.2

From the figure 7.2, we can conclude that the data has seasonality.

7.1.2.2 SEASONAL DIFFERENCING

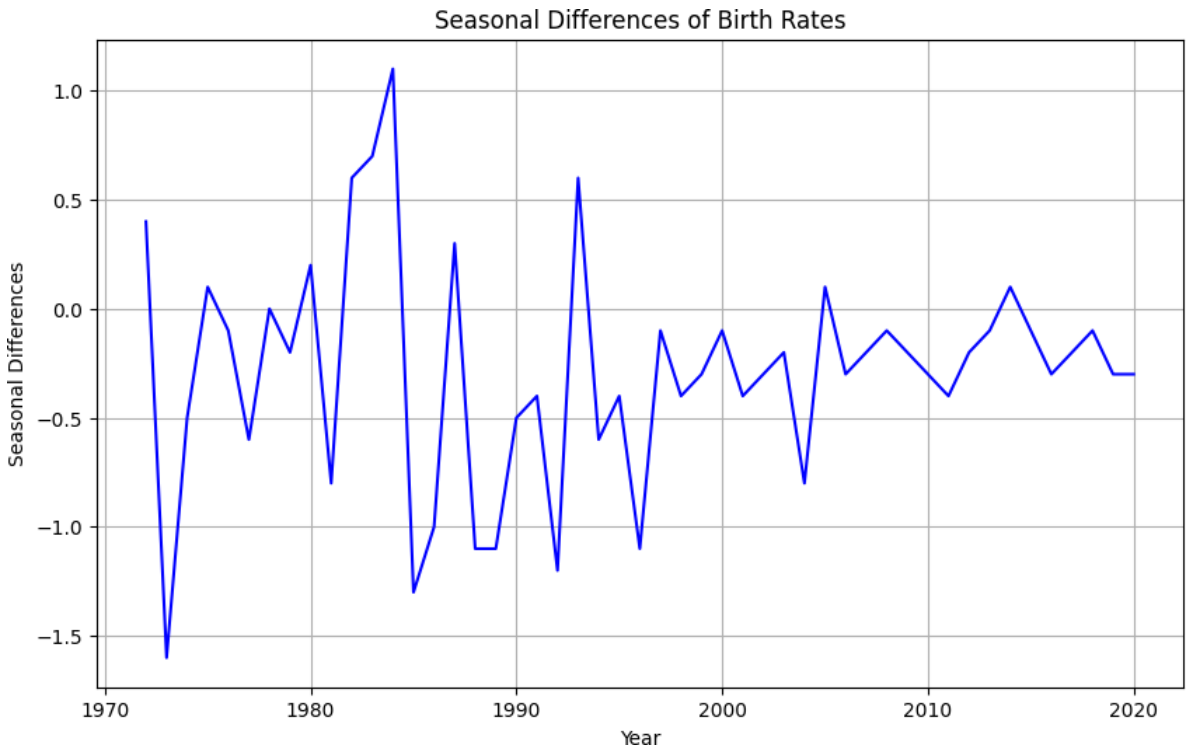


Fig 7.3

7.1.2.3 AUTOCORRELATION FUNCTION AND PARTIAL AUTOCORRELATION FUNCTION

Next we want to examine the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF). To check for stationarity, we want to plot the ACF and PACF values.

ACF and PACF plot is given in figure 7.4

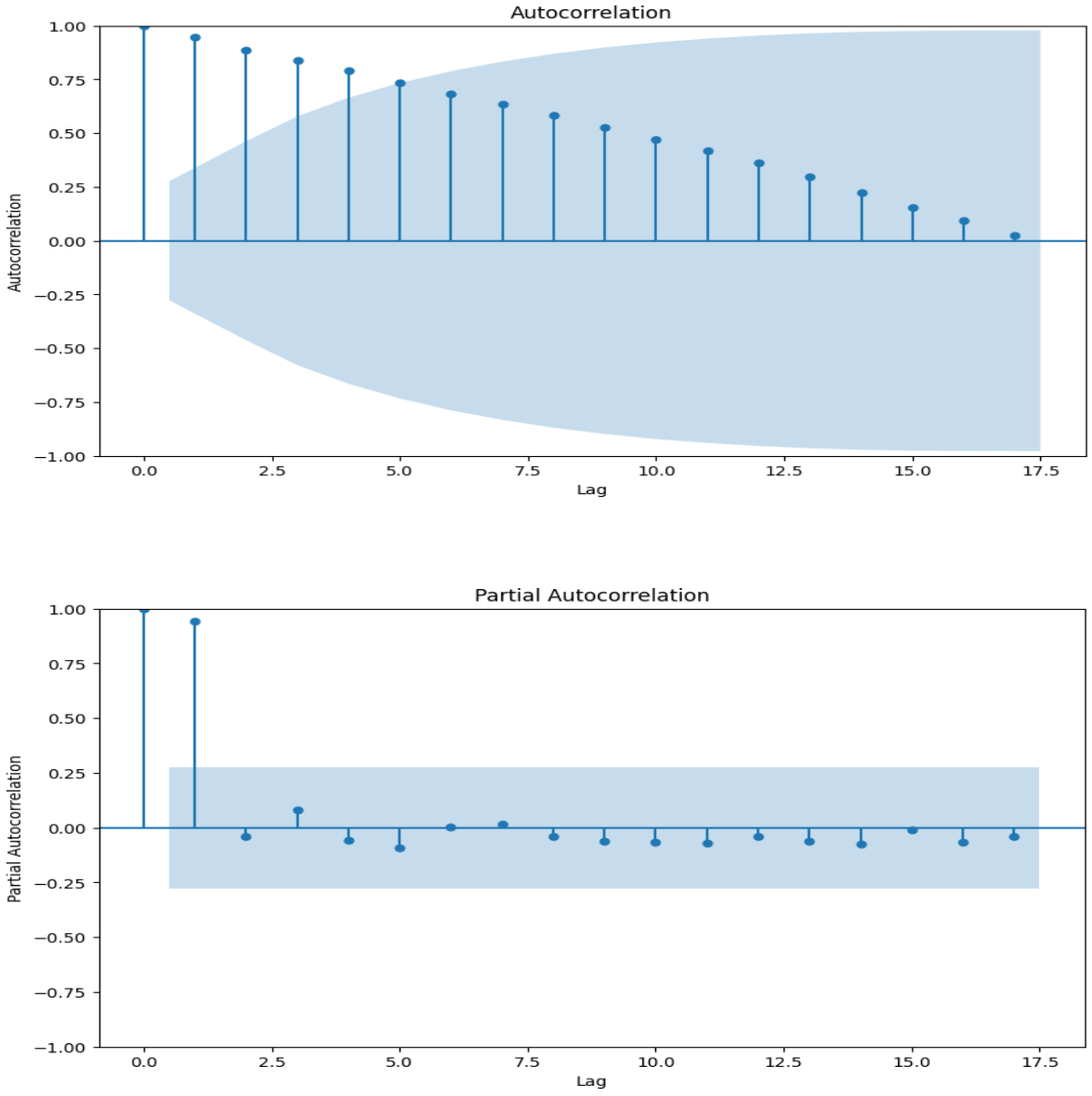


Fig 7.4

7.1.2.4 KWIATKOWSKI-PHILLIPS-SCHMIDT-SHINTEST FOR STATIONARITY

Using KPSS test, we test the time series data for stationarity whichfollows a hypothesis testing approach.

The null hypothesis H_0 is given by,

H_0 : The data isstationary

The alternative hypothesis is given by,

H_1 : The data is non stationary

The outcome obtained is, table 7.3

KPSSstatistic	1.0878607387874208
Lagorder	4
P–value	0.01

Table 7.3

The KPSS test gives the p-value 0.01, it is less than 0.05.

So we fail to rejectH_0 and hence we can conclude that data is non-stationary.

7.1.2.5 ROLLING STATISTICS PLOT

Also we can plot the rolling statistics plot to check for stationarity. The rolling mean and standard deviation plot is given in figure 7.5

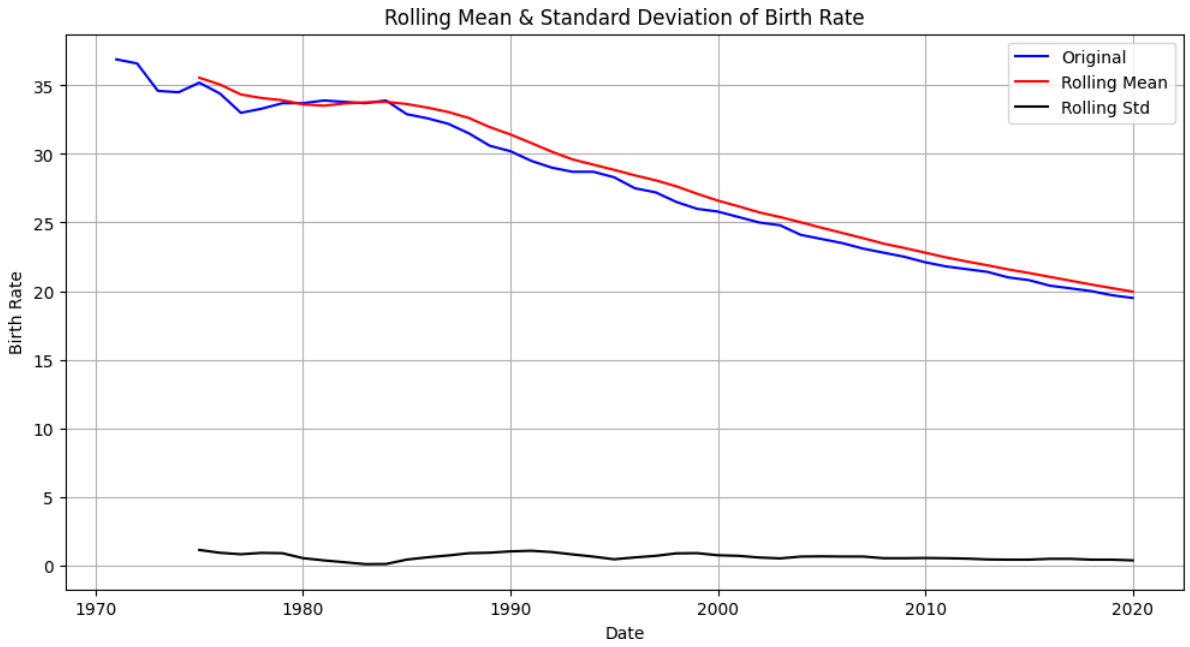


Fig 7.5

From the figure, we can observe that the data is non-stationary.

7.1.2.6 SARIMA MODEL FOR BIRTH RATE

Next step is to choose the best model for forecasting. Choose the best model from all possible models according to Akaike Information Criterion (AIC). The model with lowest AIC value is the best model. Thus, the possible time series models along with their corresponding AIC statistic for the natural logarithm of birth rates data are shown in table 7.4

SL.NO	MODEL ARIMA(p, d, q)x(P,D,Q)	AIC
1.	ARIMA(0,1,0)x(0,1,1) [4]	71.96455057083836
2.	ARIMA(0, 1, 0)x(1, 0, 0) [4]	71.92621995198992
3.	ARIMA (0, 1, 0)x(1, 1, 1) [4]	73.8573266451284
4.	ARIMA (1, 0, 0)x(1, 0, 1) [4]	67.04604589986062
5.	ARIMA (1, 0, 1)x(1, 0, 0) [4]	69.15138063156628
6.	ARIMA (1, 0, 1)x(1, 0, 1) [4]	68.71286140111155
7.	ARIMA (1, 0, 0)x(0, 0, 1) [4]	67.43222177733526
8.	ARIMA (0, 1, 1)x(1, 0, 0) [4]	72.70014997863728
9.	ARIMA (0, 1, 1)x(1, 0, 1) [4]	72.5336312092503
10.	ARIMA (0, 1, 1)x(1, 1, 1) [4]	73.72373683167801

Table 7.4

Here the best model is ARIMA (1,0,0) x (1,0,1) [4] with AIC value 67.04604589986062.

Coefficients: table 7.5

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.0196	0.157	0.124	0.901	-0.289	0.328
ar.S.L4	0.2131	0.207	1.027	0.304	-0.194	0.620
ma.S.L4	-1.0000	2782.097	-0.000	1.000	-5453.809	5451.809
sigma2	0.2553	710.303	0.000	1.000	-1391.913	1392.423

Table 7.5

7.1.2.7 DIAGNOSTIC CHECKING

Diagnostic checking is essential for ensuring the reliability, validity, and effectiveness of statistical models. It helps to choose the right model, estimate parameters correctly and improve prediction accuracy.

Diagnostic plot is given in fig.7.6

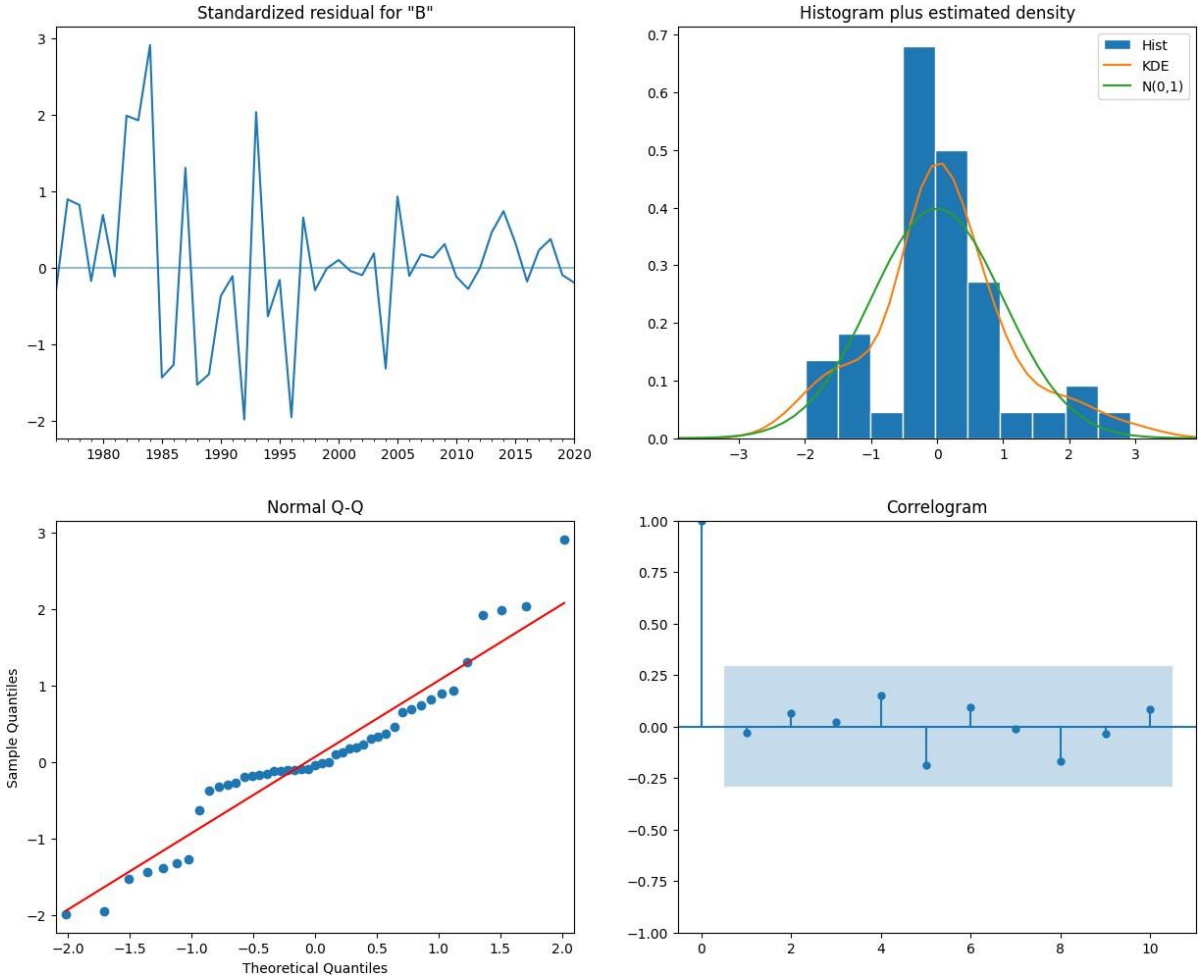


Fig 7.6

From the Q-Q plot, most of the residuals are located on the straight line and so the standard.

Residuals of fitted model seems to be normal.

7.1.2.8 IN SAMPLE FORECAST

In sample forecast refers to a prediction made by a model that uses data points within the range of data it was trained on. The table 7.6 is the actual and in sample forecasted values and fig 7.6 is the plot of actual and predicted values.

Date	Actual value	Predicted value
1976-01-01	28.4	28.575167
1977-01-01	27.8	27.306863
1978-01-01	27.8	27.347157
1979-01-01	27.6	27.694814
1980-01-01	27.8	27.463162
…	…	…
2016-01-01	17.0	17.080959
2017-01-01	16.8	16.696130
2018-01-01	16.7	16.529076
2019-01-01	16.4	16.442978
2020-01-01	16.1	16.187897

Table 7.6

Plot of actual values vs predicted values

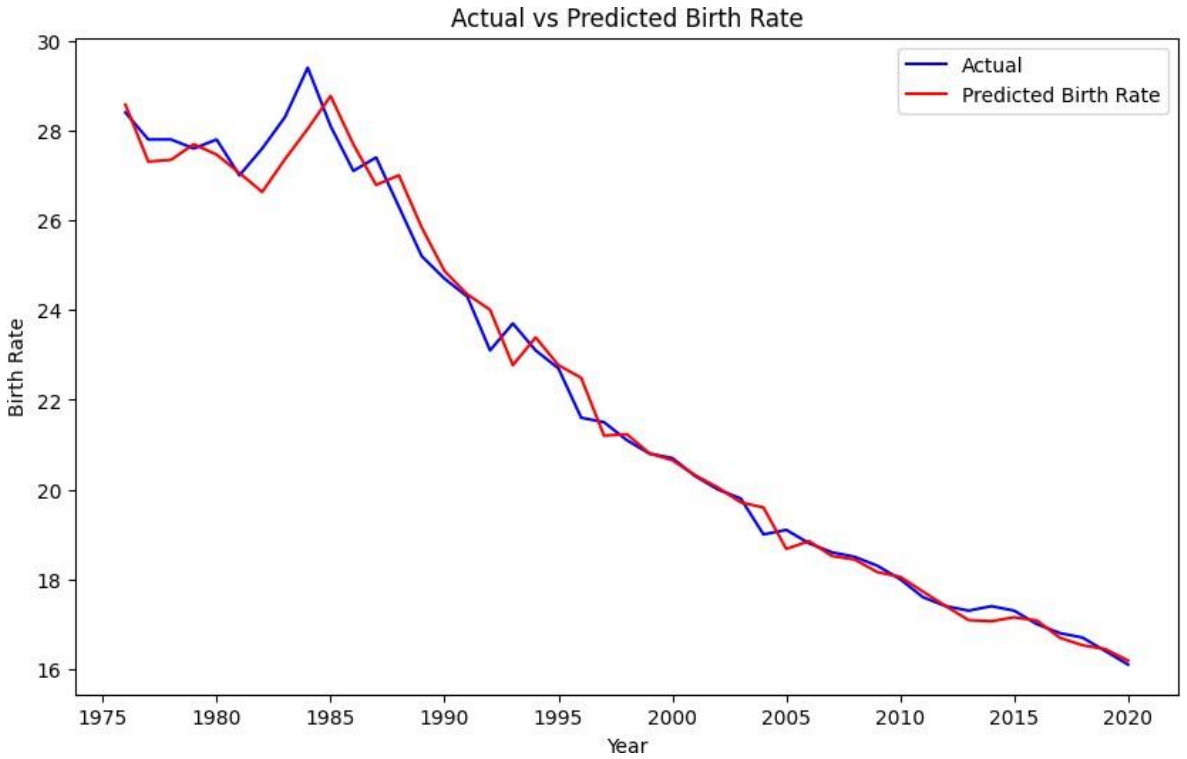


Fig 7.7

7.1.2.9 FORECASTING

The forecast values of birth rates during 2021 to 2030 is given in the table 7.7

Date	Future prediction
2021-01-01	15.820930
2022-01-01	15.556655
2023-01-01	15.325587
2024-01-01	15.126310
2025-01-01	14.879036
2026-01-01	14.641669
2027-01-01	14.424993
2028-01-01	14.228160
2029-01-01	14.004346
2030-01-01	13.787502

Table 7.7

Here is the UCL and LCL

Date	Lower Birth Rate	Upper Birth Rate
1971-01-01	-1959.963985	1959.963985
1972-01-01	-1930.306897	1989.621475
1973-01-01	-1929.912780	1990.015592
1974-01-01	-1931.489247	1988.439124
1975-01-01	-1914.324826	2005.603545
…	…	…
2016-01-01	16.192773	17.969146
2017-01-01	15.807953	17.584326
2018-01-01	15.640890	17.417263
2019-01-01	15.554791	17.331164
2020-01-01	15.299745	17.076049

Table 7.8

The graphical representation of the forecasted values of birth rates in urban area is shown in fig 7.8.

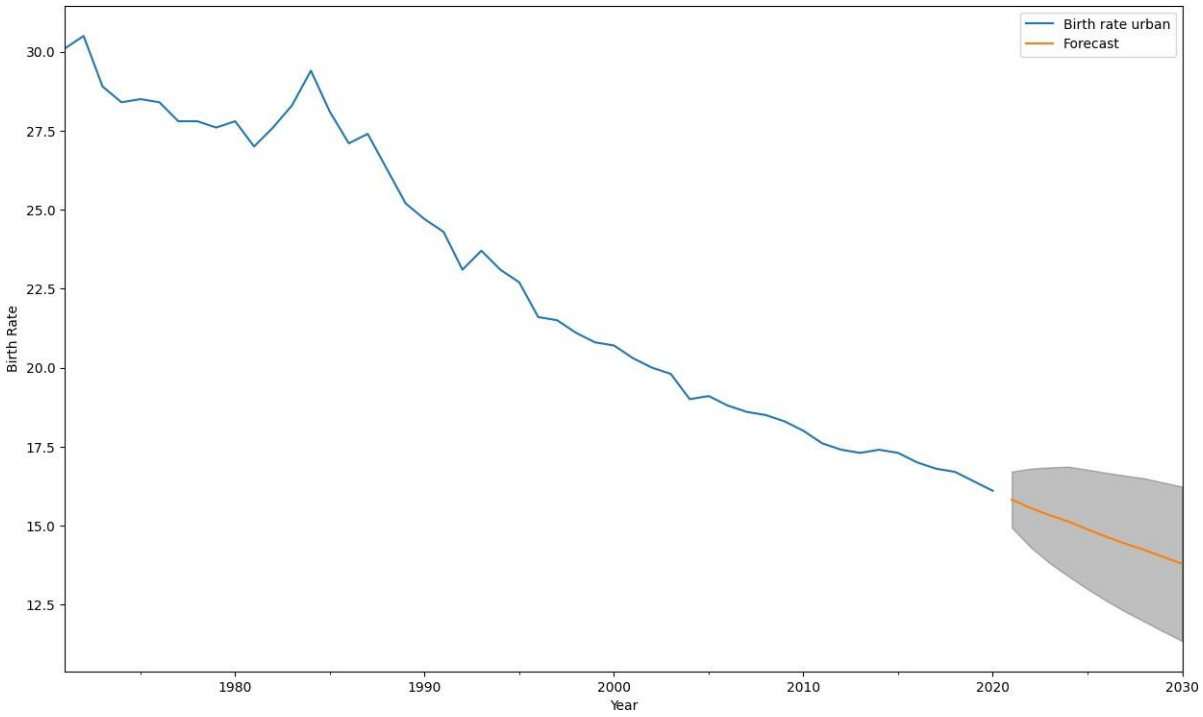


Fig 7.8

7.1.3 RECURRENT NEURAL NETWORK RNN

The given forecasted values of birth rates for the period from 2021 to 2030 and the forecasting technique used to generate these forecasting values is simple RNN. RNN used to predict the future trends based on historical data. The forecast values of Birth rates during 2021to 2030 is given in the following table 7.9.

Future Dates	Predicted values
2021-01-01	15.709132
2022-01-01	15.434744
2023-01-01	15.004195
2024-01-01	14.675736
2025-01-01	14.238377
2026-01-01	13.884169
2027-01-01	13.413366
2028-01-01	13.036413
2029-01-01	12.548463
2030-01-01	12.145741

Table 7.9

The trend in the forecasted values shows a consistent decrease over time. Starting from 15.709132 in 2021, the values slowly decrease in each year, reaching 12.145741 in 2030.The graphical representation of the forecast values of birth rates in urban area is shown in the fig.7.9.

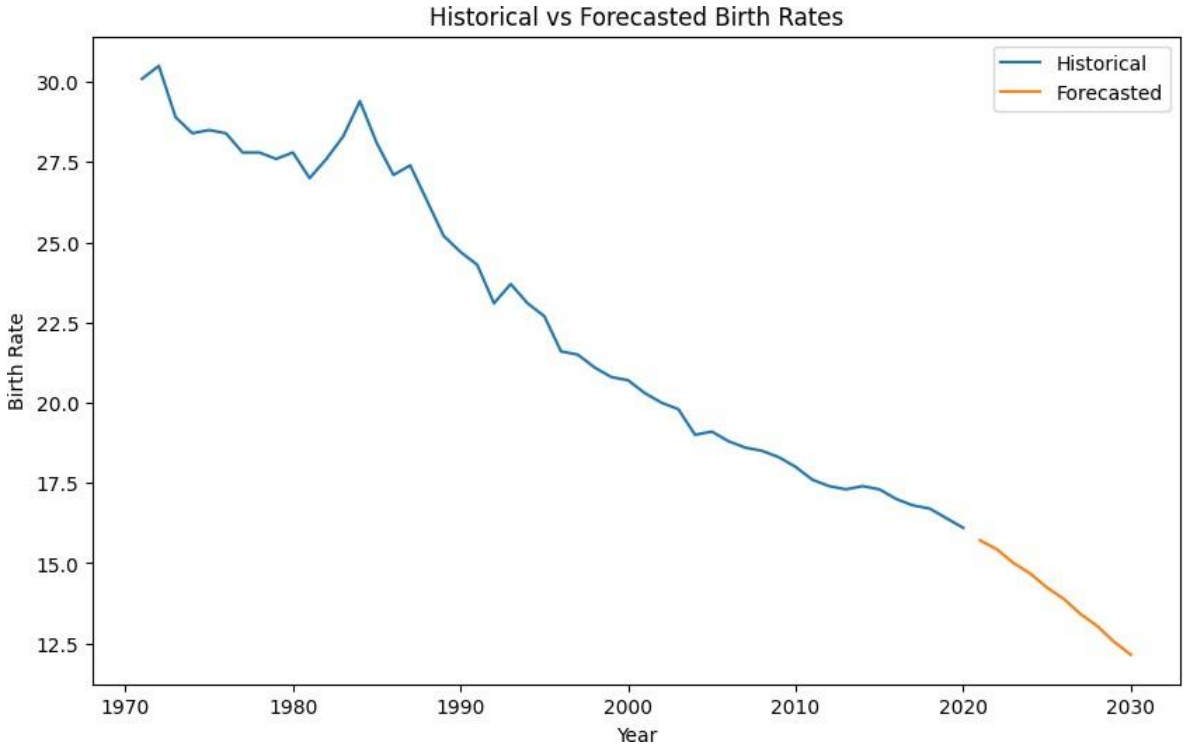


Fig 7.9

7.1.4 COMPARE MSE & RMSE VALUES

To determine which model is the best, the performance metrics should be compared andconsider the context of the problem being addressed. In this case, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were provided for both the RNN model and SARIMA model.Comparing MSE and RMSE values of SARIMA model and RNN model is given in the following table 7.10.

	SARIMA MODEL	**RNN MODEL**
MSE	**0.21977544148558362**	**3.07528115869258**
RMSE	**0.4688021346853954**	**9.457354205009576**

Table 7.10

The SARIMA model has a lower MSE and RMSE compared to the RNN model. Based on the provided metrics alone, the SARIMA model appears to perform better than the RNN model in terms of MSE and RMSE.

7.2 RURAL BIRTH RATE

Table 7.11 shows that there is no null value in the given data

Birth rate rural	**0**
dtype	**Int64**

Table 7.11

Table 7.12 shows the data summary

	Birth rate rural
Minimum	**21.100000**
1st Quartile	**24.475000**
Mean	**29.436000**
Median	**29.650000**
S D	**5.369481**
3rd Quartile	**34.600000**
Maximum	**38.900000**

Table 7.12

The time series plot of birth rates in India is given in figure 7.10.

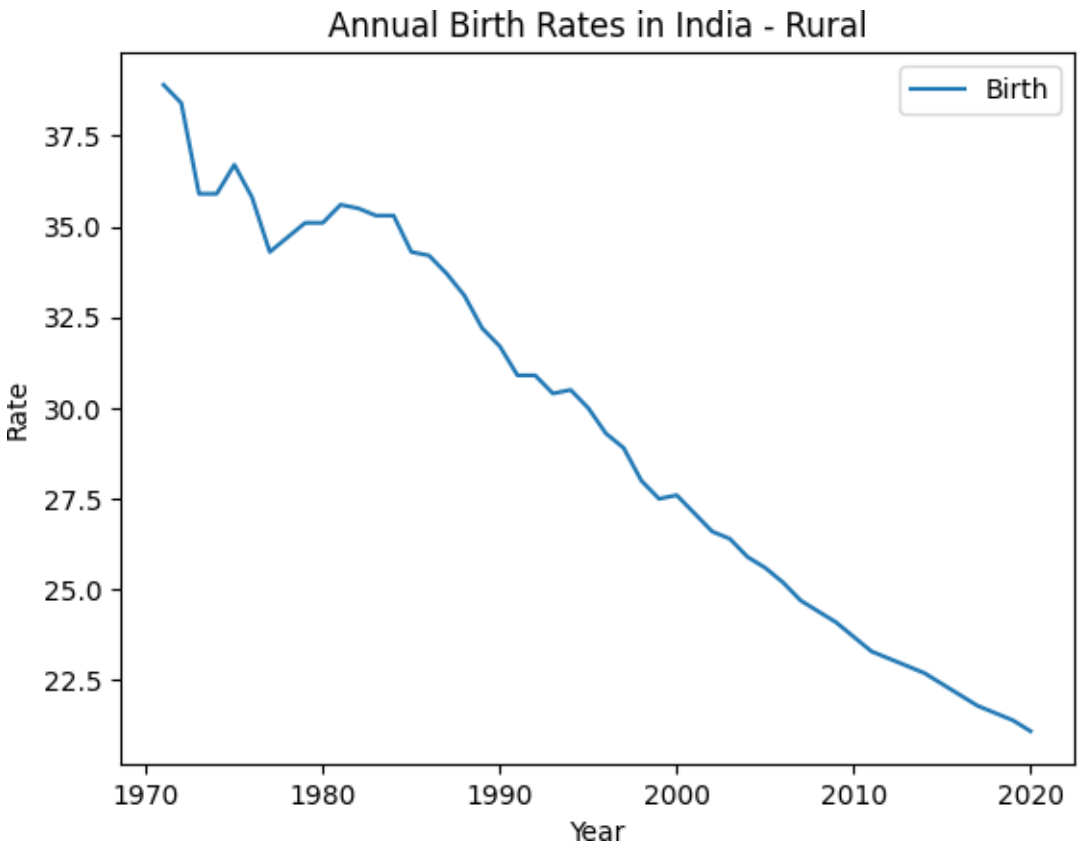


Fig 7.10

7.2.2 TIME SERIES

7.2.2.1 DECOMPOSITION OF TIME

Do seasonal decomposition for evaluating the trend, seasonal and random components of the time series and draw the seasonal plot. The seasonal plot is given in figure 7.11.

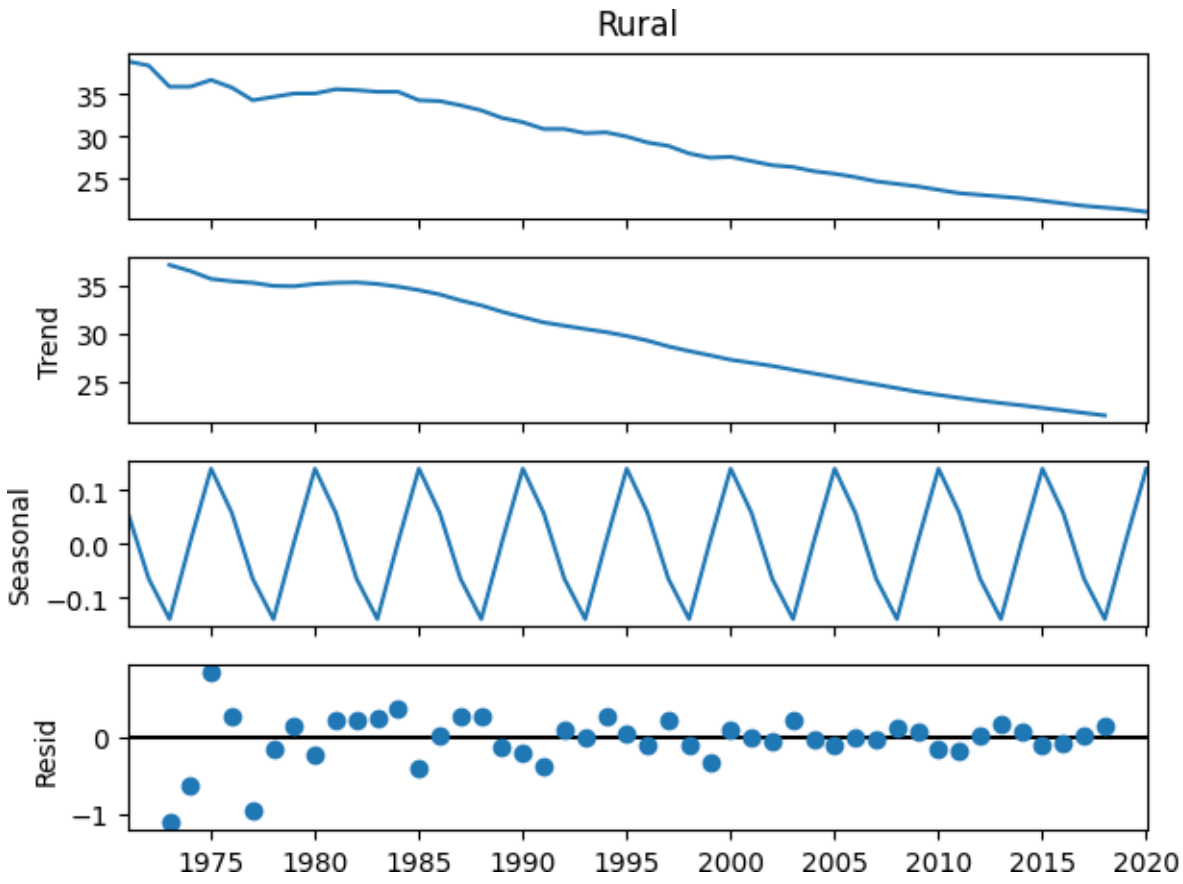


Fig 7.11

From the figure 7.11, we can conclude that the data has seasonality.

7.2.2.2 SEASONAL DIFFERENCING

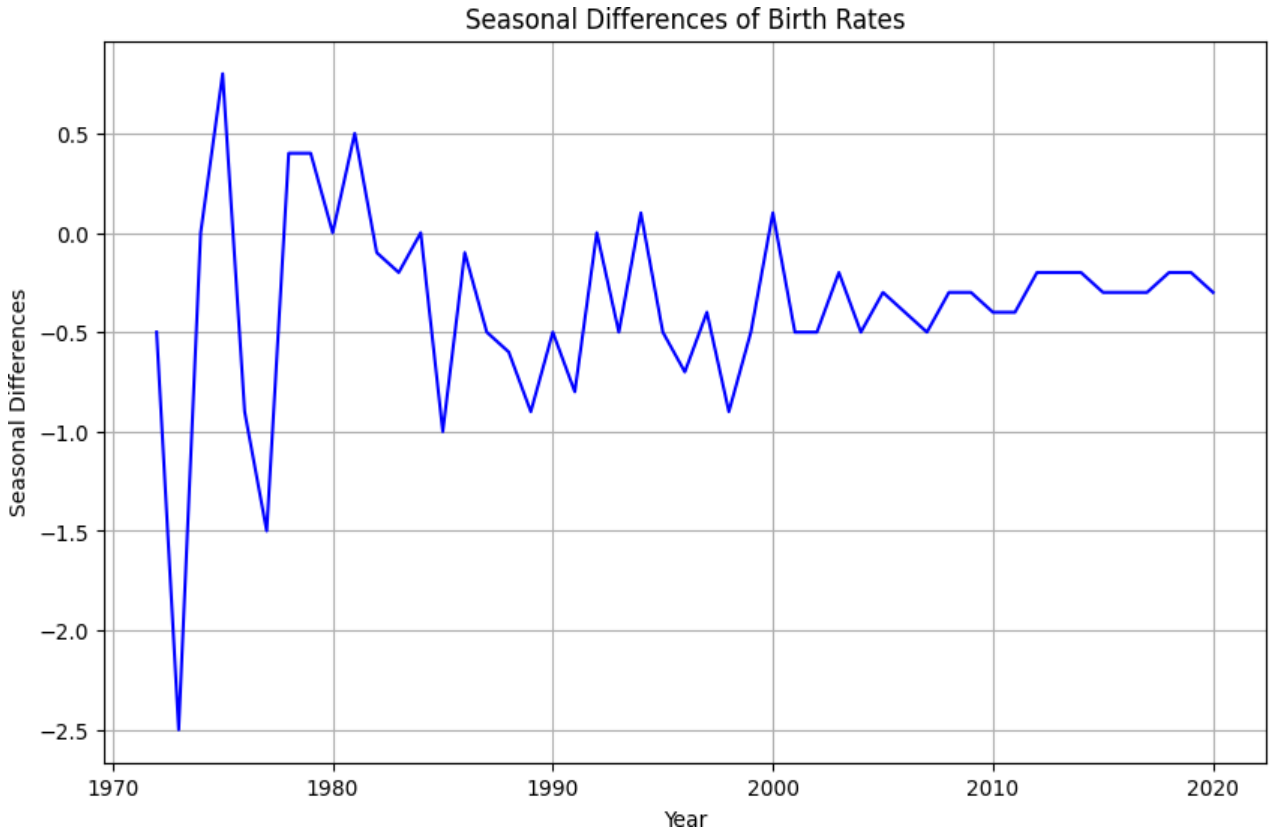


Fig 7.12

7.2.2.3 AUTOCORRELATION FUNCTION AND PARTIAL AUTOCORRELATION FUNCTION

Next we want to examine the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF). To check for stationarity, we want to plot the ACF and PACF values.

ACF and PACF plot is given in figure 7.13

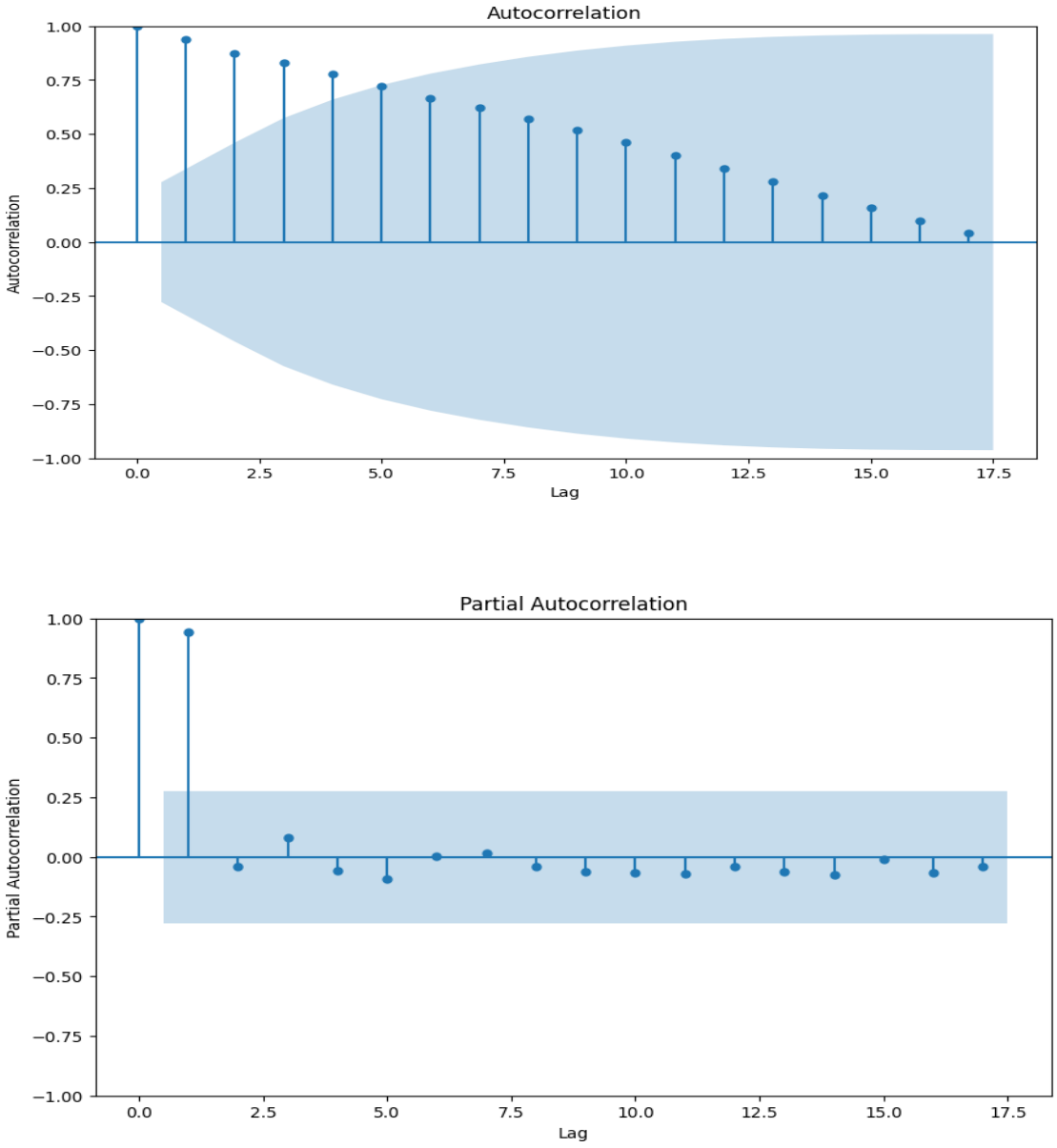


Fig 7.13

7.2.2.4 KWIATKOWSKI-PHILLIPS-SCHMIDT-SHINTEST FOR STATIONARITY

Using KPSS test, we test the time series data for stationarity which follows a hypothesis testing approach.

The null hypothesis H_0 is given by,

H_0 : The data is stationary

The alternative hypothesis is given by,

H_1 : The data is non stationary

The outcome obtained is, table 7.13

KPSSstatistic	1.104413773769745
Lagorder	4
P–value	0.01

Table 7.13

The KPSS test gives the p-value 0.01, it is less than 0.05.

So we fail to reject H_0 and hence we can conclude that data is non-stationary.

7.2.2.5 ROLLING STATISTICS PLOT

Also we can plot the rolling statistics plot to check for stationarity. The rolling mean and standard deviation plot is given in figure 7.14

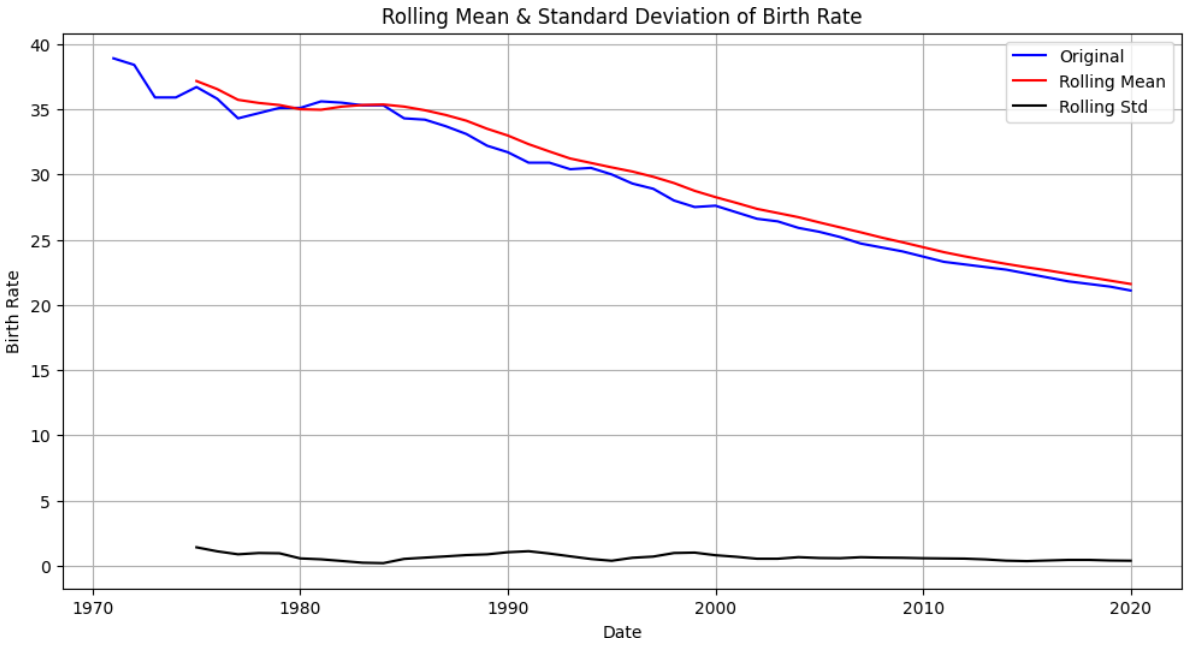


Fig 7.14

From the figure, we can observe that the data is non-stationary.

7.2.2.6 SARIMA MODEL FOR BIRTH RATE

Next step is to choose the best model for forecasting. Choose the best model from all possible models according to Akaike Information Criterion (AIC). The model with lowest AIC value is the best model. Thus, the possible time series models along with their corresponding AIC statistic for the natural logarithm of birth rates data are shown in table 7.14

SL.NO	MODEL ARIMA(p, d, q)x(P,D,Q)	AIC
1.	ARIMA(1,1,1)x(1,1,1) [4]	32.84185607294047
2.	ARIMA(1, 1, 1)x(1, 0, 1) [4]	36.5956090815806
3.	ARIMA (1, 1, 1)x(0, 1, 1) [4]	33.20153554398637
4.	ARIMA (1, 0, 1)x(1, 0, 1) [4]	33.52509717402558
5.	ARIMA (1, 0, 0)x(1, 0, 1) [4]	31.45595470886906
6.	ARIMA (0, 1, 1)x(0, 1, 1) [4]	30.7589836146784
7.	ARIMA (1, 0, 1)x(0, 0, 1) [4]	45.1890113736854
8.	ARIMA (1, 0, 1)x(1, 0, 0) [4]	42.34230962505152
9.	ARIMA (1, 1, 1)x(0, 0, 1) [4]	45.83084300729781
10.	ARIMA (0, 1, 0)x(1, 1, 1) [4]	46.38426338614562

Table 7.14

Here the best model is ARIMA (0,1,1) x (0,1,1) [4] with AIC value30.7589836146784.

Coefficients: table 7.15

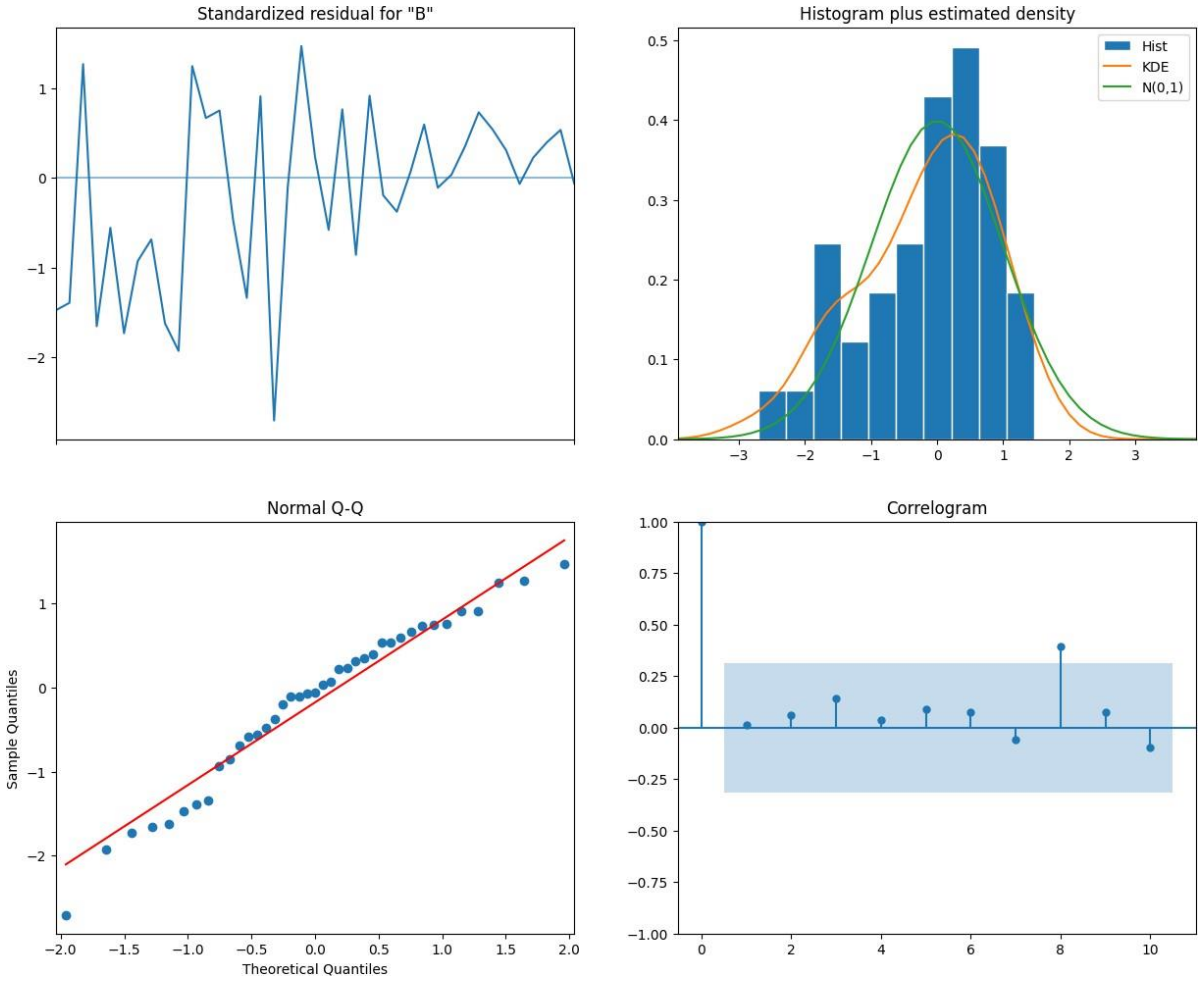
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.0101	0.105	0.096	0.923	-0.196	0.216
ar.S.L4	-0.1361	0.152	-0.896	0.370	-0.434	0.162
ma.S.L4	--0.5355	0.255	-2.097	0.036	-1.036	-0.035
sigma2	0.1021	0.025	4.019	0.000	0.052	0.152

Table 7.15

7.2.2.7 DIAGNOSTIC CHECKING

Diagnostic checking is essential for ensuring the reliability, validity, and effectiveness of statistical models. It helps to choose the right model, estimate parameters correctly and improve prediction accuracy.

Diagnostic plot is given in fig.7.15



7.15

From the Q-Q plot, most of the residuals are located on the straight line and so the standard.

Residuals of fitted model seems to be normal.

7.2.2.8 IN SAMPLE FORECAST

In sample forecast refers to a prediction made by a model that uses data points within the range of data it was trained on. The table 7.15 is the actual and in sample forecasted values and fig 7.16 is the plot of actual and predicted values.

Date	Actual value	Predicted value
1976-01-01	35.8	36.234036
1977-01-01	34.3	33.470179
1978-01-01	34.7	34.300000
1979-01-01	35.1	37.677258
1980-01-01	35.6	34.254457
…	…	…
2016-01-01	22.1	22.122457
2017-01-01	21.8	21.790199
2018-01-01	21.6	21.450557
2019-01-01	21.4	21.225618
2020-01-01	21.1	21.126788

Table 7.16

Plot of actual values vs predicted values

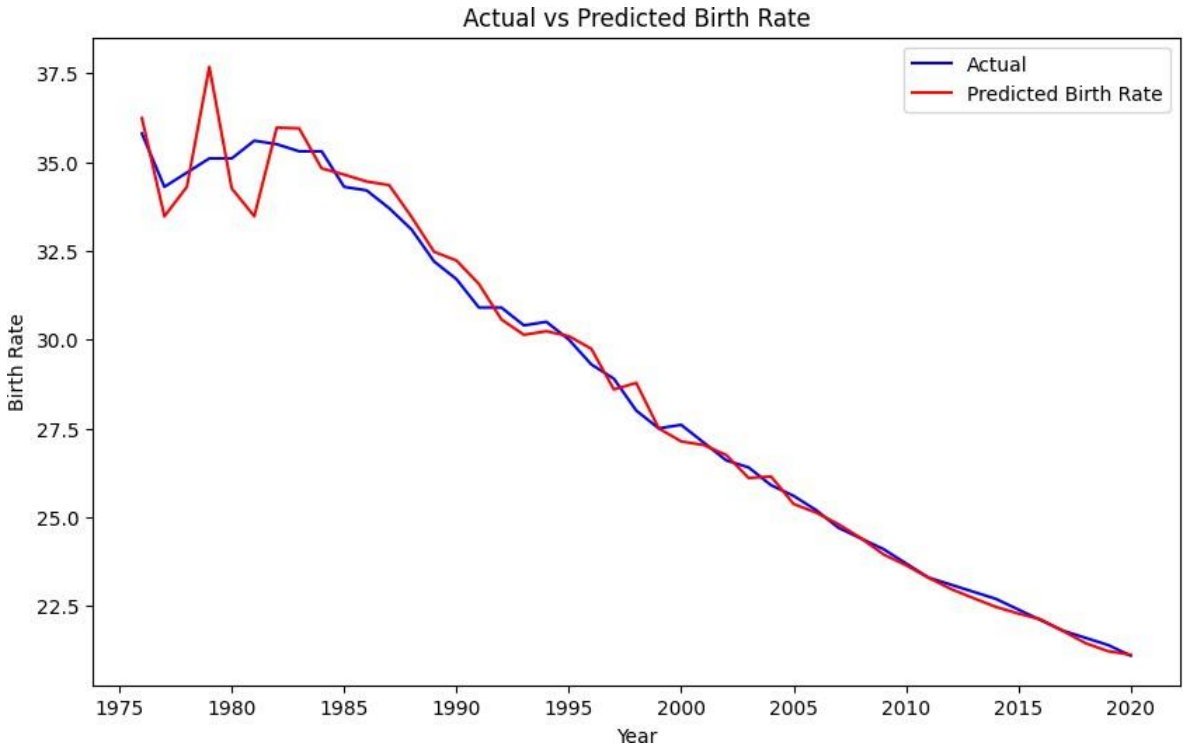


Fig 7.16

7.2.2.9 FORECASTING

The forecast values of birth rates during 2021 to 2030 is given in the table.

Date	Future prediction
2021-01-01	**20.808216**
2022-01-01	**20.528138**
2023-01-01	**20.220333**
2024-01-01	**19.933732**
2025-01-01	**19.640975**
2026-01-01	**19.371799**
2027-01-01	**19.078671**
2028-01-01	**18.790246**
2029-01-01	**18.497621**
2030-01-01	**18.226961**

Table 7.17

Here is the UCL and LCL

Date	Lower Birth Rate	Upper Birth Rate
1971-01-01	**-3394.757202**	**3394.757202**
1972-01-01	**-2732.907719**	**2810.707719**
1973-01-01	**-2733.407719**	**2810.207719**
1974-01-01	**-2410.000826**	**2807.707719**
1975-01-01	**-1932.791128**	**2530.136842**
…	**…**	**…**
2016-01-01	**21.496270**	**22.748644**
2017-01-01	**21.164012**	**22.416386**
2018-01-01	**20.824370**	**22.076744**
2019-01-01	**20.599431**	**21.851805**
2020-01-01	**20.500602**	**21.752974**

Table 7.18

The graphical representation of the forecasted values of birth rates in urban area is shown in fig 7.8.

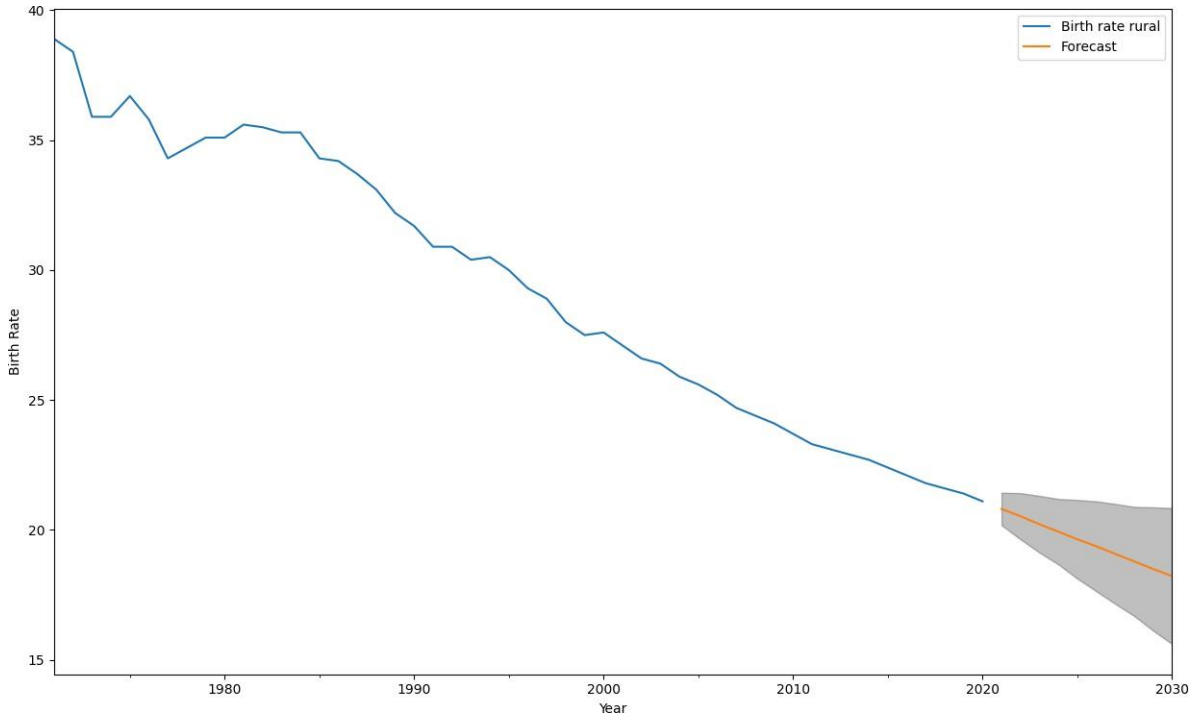


Fig 7.17

7.2.3 RECURRENT NEURAL NETWORK RNN

The given forecasted values of birth rates for the period from 2021 to 2030 and the forecasting technique used to generate these forecasting values is simple RNN. RNN used to predict the future trends based on historical data. The forecast values of Birth rates during 2021to 2030 is given in the following table 7.18.

Future Dates	Predicted values
2021-01-01	**20.805626**
2022-01-01	**20.551170**
2023-01-01	**20.282444**
2024-01-01	**19.997501**
2025-01-01	**19.733517**
2026-01-01	**19.482691**
2027-01-01	**19.230114**
2028-01-01	**18.983986**
2029-01-01	**19.752014**
2030-01-01	**19.527279**

Table 7.19

The trend in the forecasted values shows a consistent decrease over time. Starting from 20.805626 in 2021, the values slowly decrease in each year, reaching 19.527279 in 2030. The graphical representation of the forecast values of birth rates in ruralarea is shown in the fig.7.9.

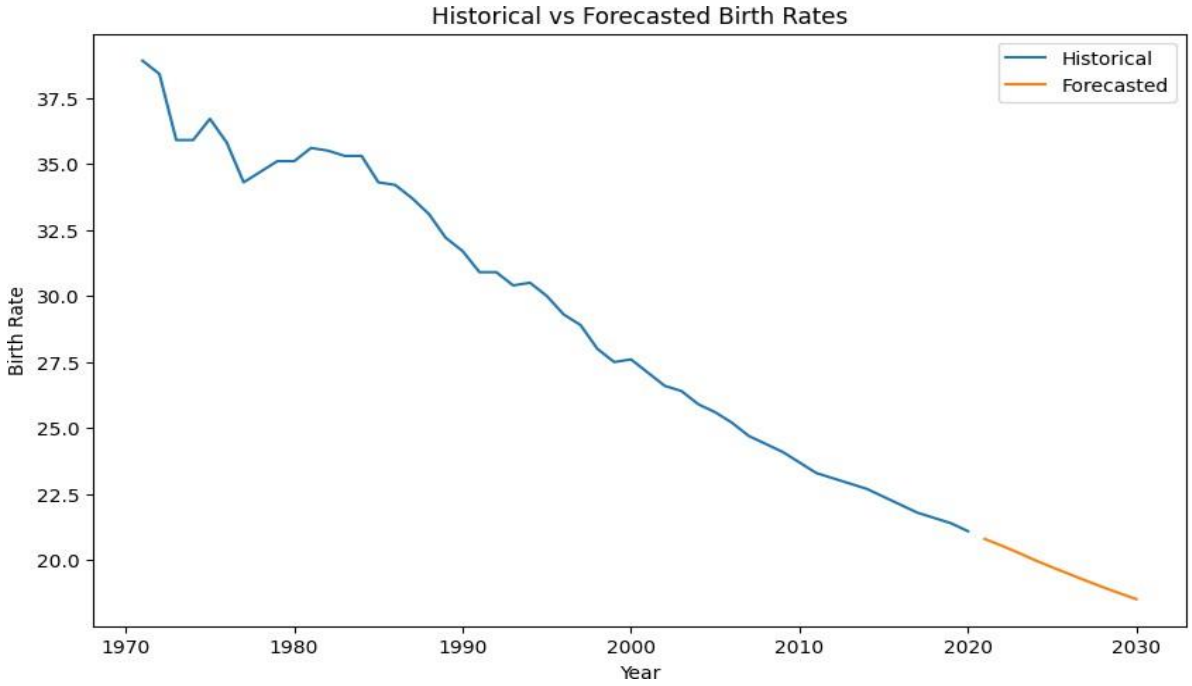


Fig 7.18

7.2.4 COMPARE MSE & RMSE VALUES

To determine which model is the best, the performance metrics should be compared and consider the context of the problem being addressed. In this case, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were provided for both the RNN model and SARIMA model. Comparing MSE and RMSE values of SARIMA model and RNN model is given in the following table 7.11.

	SARIMA MODEL	**RNN MODEL**
MSE	**0.38025391918453916**	**2.6060034591233756**
RMSE	**0.6166473215579057**	**6.791254028962999**

Table 7.20

The SARIMA model has a lower MSE and RMSE compared to the RNN model. Based on the provided metrics alone, the SARIMA model appears to perform better than the RNN model in terms of MSE and RMSE.

CHAPTER – 8
CONCLUSION

Based on the results, after analysing the performance of SARIMA and simple RNN models for forecasting birth rates in urban and rural areas based on data from 1971 to 2020, the SARIMA model appears to be more effective in predicting the birth rates. The accuracy of these forecasted values depends on the quality of data, regular updates are essential for better predictions.

There is growing concern in this study that forecast birth rates from 1971 to 2020 shows further decline. It indicates some seasonal changes, which refer to changes in birth rates and design that occur periodically throughout the year due to factors such as urbanisation, education and economy.

For Urban areas, seasonal changes are related to economic opportunities, family planning policies, cost of living, education and empowerment of women, social and cultural norms and quality of life.

For Urban areas, seasonal changes are related to economic changes, education and employment opportunities, access to health care and family planning services, migration to urban areas, social and cultural changes and Government policies.

In summary, forecast patterns in birth rates will vary with different areas. Forecasting helps make informed decisions and provide valuable insights for policymakers, healthcare providers, educators, businesses, and community organizations to plan and respond effectively to demographic changes and promote the health, well-being, and prosperity of populations.

CHAPTER – 9
REFERENCES

1. Baghestani, H., & Malcolm, M. (2016). Factors predicting the US birth rate. Journal of Economic Studies, 43(3), 432-446.
2. Chichkanov, V. P., Vasilyeva, A. V., Bystray, G. P., &Okhotnikov, S. A. (2015). Russia's Birth Rate Dynamics Forecasting. R-Economy, 1(2), 351-356.
3. Ermisch, J. (1988). Econometric analysis of birth rate dynamics in Britain. The Journal of Human Resources, 23(4), 563-576.
4. Heckman, J. J., & Walker, J. R. (1989). Forecasting aggregate period-specific birth rates: the time series properties of a microdynamic neoclassical model of fertility. Journal of the American Statistical Association, 84(408), 958-965.
5. Karunanidhi, D., & Sasikala, S. (2023). Robustness of Predictive Performance of Arima Models Using Birth Rate of Tamilnadu. Journal of Statistics Applications and Probability, 12(3), 1189-201.
6. Kim, K. W., Li, G., Park, S. T., & Ko, M. H. (2016). A study on birth prediction and bcg vaccine demand prediction using arima analysis. Indian Journal of Science and Technology.
7. Lee, R. D. (1993). Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level. International Journal of forecasting, 9(2), 187-202.
8. Leung, C. C. S. (1995). Time series modelling of birth data.
9. McDonald, J. (1979). A time series approach to forecasting Australian total live-births. Demography, 16, 575-601.
10. Montgomery, D. C., Jennings, C. L., &Kulahci, M. (2015). Introduction to time series analysis and forecasting. John Wiley & Sons.