FORECASTING GLOBAL DEMOGRAPHICS:

TIME SERIES ANALYSIS OF POPULATION GROWTH

A PROJECT REPORT

Submitted by

CB.SC.I5MAT21004 **GAYATHRI S**

Under the guidance of

Dr. Diya Bhattacharyya

in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF SCIENCE

IN

MATHEMATICS

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE 641112

MAY 2024

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112

BONAFIDE CERTIFICATE

This is to certify that the project report entitled "**FORECASTING GLOBAL DEMOGRAPHICS: TIME SERIES ANALYSIS OF POPULATION GROWTH**" submitted by **CB.SC.I5MAT21004 GAYATHRI S** in partial fulfilment of the requirements for the award of the **Degree of Bachelor of Science** in **MATHEMATICS** is a bonafide record of the work carried out under my guidance and supervision at Amrita School of Engineering, Coimbatore.

Project Advisor

Designation

Project Coordinator

Designation

Chairperson
Department of Mathematics
Dr. Somasundaram K

The project was evaluated by us on:

Internal Examiner

External Examiner

DEDICATION

To our parents and teachers

ACKNOWLEDGEMENT

My solemn gratitude to Amrita Vishwa Vidyapeetham for having provided me with the necessary preliminaries and pre-requisites for the completion of my project.

I express my deep sense of gratitude to Dr. Diya Bhattacharyya, Asst. Professor, Department of Mathematics for her guidance and encouragement throughout the duration of the project. I am indebted to her for her motivation that kept me going and helped me finish the project.

I would like to place on record, my thankfulness to the review panel members
Dr. P Tamilalagan, Dr. Gayathri K, Dr. J Geetha, Dr. K S Sreeranjini, for assessing my work.

My sincere thanks to all the other faculty members, my parents and my friends who supported me in every step of this project.

Contents

Chapter 1

INTRODUCTION TO TIME SERIES ANALYSIS

1 INTRODUCTION TO TIME SERIES

1.1 Introduction

Population analysis plays a very important role in shaping a country's future. It is crucial for effective planning and decision-making. It ensures well-being and sustainable development of the society. This dissertation concentrates on the time series analysis of increase or decrease in population of different countries and the demographic factors behind them.

Forecasting of population is important for several reasons like anticipating future needs for resources, economic development, educational planning, healthcare planning, environmental management, etc. It helps in long-term planning of various aspects of society.

With the help of time series analysis, populations of four countries have been analysed and the future population of the countries along with the population growth rate have been predicted.

The datasets used here are the population of four countries, namely, India, China, Russia and United States, from the years 1950 to 2022. Also, the use of four models of time series were explored for the best prediction. The four models used were Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA) and Autoregressive Moving Average (ARIMA) models.

It is necessary to find the best model for forecasting. There are a lot of criteria to determine the best model. The one used here is the Root Mean Square Error (RMSE) values. The model with least RMSE value gives the best result.

As forecasting is done using the best model, the result is expected to be the most accurate value. So, this can be used for decision-making and effective planning in various sectors.

1.2 Time Series

Time series analysis is a statistical technique that deals with time-ordered data. This type of data is collected and recorded at regular intervals such as hourly, daily, monthly, or yearly. The main goal of time series is to study and estimate the characteristics of a response variable concerning time which acts as the independent variable. Gross Domestic Product (GDP), stock markets, prices of gas over time, etc. are all examples of time series data.

1.3 Components of Time Series

1.3.1 Trend

The long-term movement or direction in the data is known as trend. It shows whether the data is increasing, decreasing or staying relatively constant over time.

1.3.2 Seasonality

The repeating or cyclic pattern in the data that occurs at fixed intervals is known as seasonality. For example, sale of ice-cream increases in the summer.

1.3.3 Cycle

Cycle is similar to seasonality but with more irregular intervals. These are usually influenced by economic conditions, business cycles or other external factors.

1.3.4 Noise

Random fluctuations or irregularity in the data that cannot be attributed to trend, seasonality or cycle is known as noise.

1.4 Stationary Time Series

A stationary time series is the time series whose properties are independent of the time at which the series is observed. Statistical properties such as mean, variance, etc. remain constant over time.

Two popular ways to convert non-stationary time series to stationary time series are as follows:

- **Differencing:** This technique involves computing the difference between two consecutive observations. This technique stabilizes the mean of a time series.
- **Logarithms:** This technique involves taking logarithms of the observations. Logarithms helps to stabilize the variance of a time series.

There are several methods that can be used to test for the stationarity of a time series.

1.4.1 Autocorrelation function (ACF)

It helps to access the stationarity of a time series by assessing the degree of correlation between the time series and its shifted duplicates at various time intervals. It helps to identify which lags have significant correlations, understand the patterns and properties of the time series, and then use that information to model the time series data.

1.4.2 Partial Autocorrelation Function (PACF)

The partial autocorrelation function is similar to the ACF except that it assesses correlation between two variables after adjusting for the impact of other variables. The PACF measures how a time series of records relate to one another when dependent intervening data are removed from the relationship at a previous time step.

1.4.3 Augmented Dickey-Fuller (ADF) Test

ADF Test is utilized to test for stationarity of data. It is a statistical test which is used to determine whether unit root is present in a time series dataset. If there is a unit root, it signifies the time series to be non-stationary. Under the null hypothesis, the time series is assumed to be non-stationary. The null hypothesis is rejected when the p-value is less than a pre-determined value, which is usually set at 0.05.

1.5 Time Series Models

Models of time series data can have many forms and represent different stochastic processes.

1.5.1 Autoregressive (AR) Model

AR model uses past values for forecasting future data. An autoregressive model of order p [AR(p)] can be written as follows:

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + e_t \quad (1.1)$$

where e_t is white noise,

y_t denotes the value of the variable at time t, and

φ_i's are the model parameters.

1.5.2 Moving Average (MA) Model

MA model uses past errors for forecasting future data. A moving average model of order q [MA(q)] can be written as follows:

$$y_t=c+e_t+\theta_1 e_{t-1}+\theta_2 e_{t-2}+\ldots+\theta_q e_{t-q} \quad (1.2)$$

where e_t is white noise,

y_t denotes the value of the variable at time t, and

θ_i's are the model parameters.

1.5.3 Autoregressive Moving Average (ARMA) Model

ARMA model is a combination of both autoregressive and moving average models, and uses both past values and past errors for predicting future data.

1.5.4 Autoregressive Integrated Moving Average (ARIMA) Model

A non-seasonal ARIMA model is obtained by combining differencing with autoregressive and moving average models. ARIMA (p, d, q) model can be written as follows:

$$y_t=c+\varphi_1 y'_{t-1}+\varphi_2 y'_{t-2}+\ldots+\varphi_p y'_{t-p}+\theta_1 e_{t-1}+\theta_2 e_{t-2}+\ldots+\theta_q e_{t-q}+e_t \quad (1.3)$$

where p=order of the autoregressive part,

d=degree of first differencing involved,

q=order of the moving average part.

1.6 Assessment of the best model

Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It measures how far predictions fall from true values using Euclidean distance and is defined as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}$$

where y_i and $\hat{y}_i$ denote the true and the observed values respectively and n is the number of observations.

Chapter 2

ANALYSING AND FORECASTING POPULATION USING TIME SERIES MODELS

2 ANALYSING AND FORECASTING POPULATION USING TIME SERIES MODELS

2.1 Introduction

Population of four countries, namely, India, China, Russia and United States of America, has been used for the study. In order to find the best model for forecasting, the data sets have been divided into two parts - training set comprising data from 1950 to 2017, on which the models will be fit, and testing set containing data from 2018 to 2022, on which the fitted models will be validated. Four models have been used, namely, AR, MA, ARMA and ARIMA. The model which has been found to have the least RMSE has been used to forecast the population of the countries in 2025. The rate of population change has also been calculated. Finally, the demographic factors that may possibly influence such changes have been discussed.

2.2 Description of data sets

The data sets used in the study are open-source data which can be easily accessed on the internet. The data sets consist of the population of four countries, India, China, Russia and United States, from the year 1950 to 2022.

2.3 Stationarity of data

Since most time series models can only be used on stationary data, the stationarity of the data sets is first checked using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots as well as Augmented Dickey-Fuller (ADF) Test.

The results of ADF test have been presented in Table 2.1. Since all p-values are greater than 0.05, we cannot reject the null-hypothesis of non-stationarity. Thus, the data are non-stationary and must be made stationary via appropriate methods.

Country	p-value
India	0.9712
China	0.99
Russia	0.7319
United States	0.5033

Table 2.1: p-values obtained by ADF test for the data sets

Figures (2.1) and (2.2) present the ACF and the PACF plots respectively for the four countries. From the figures, it may also be seen that the data sets are not stationary.

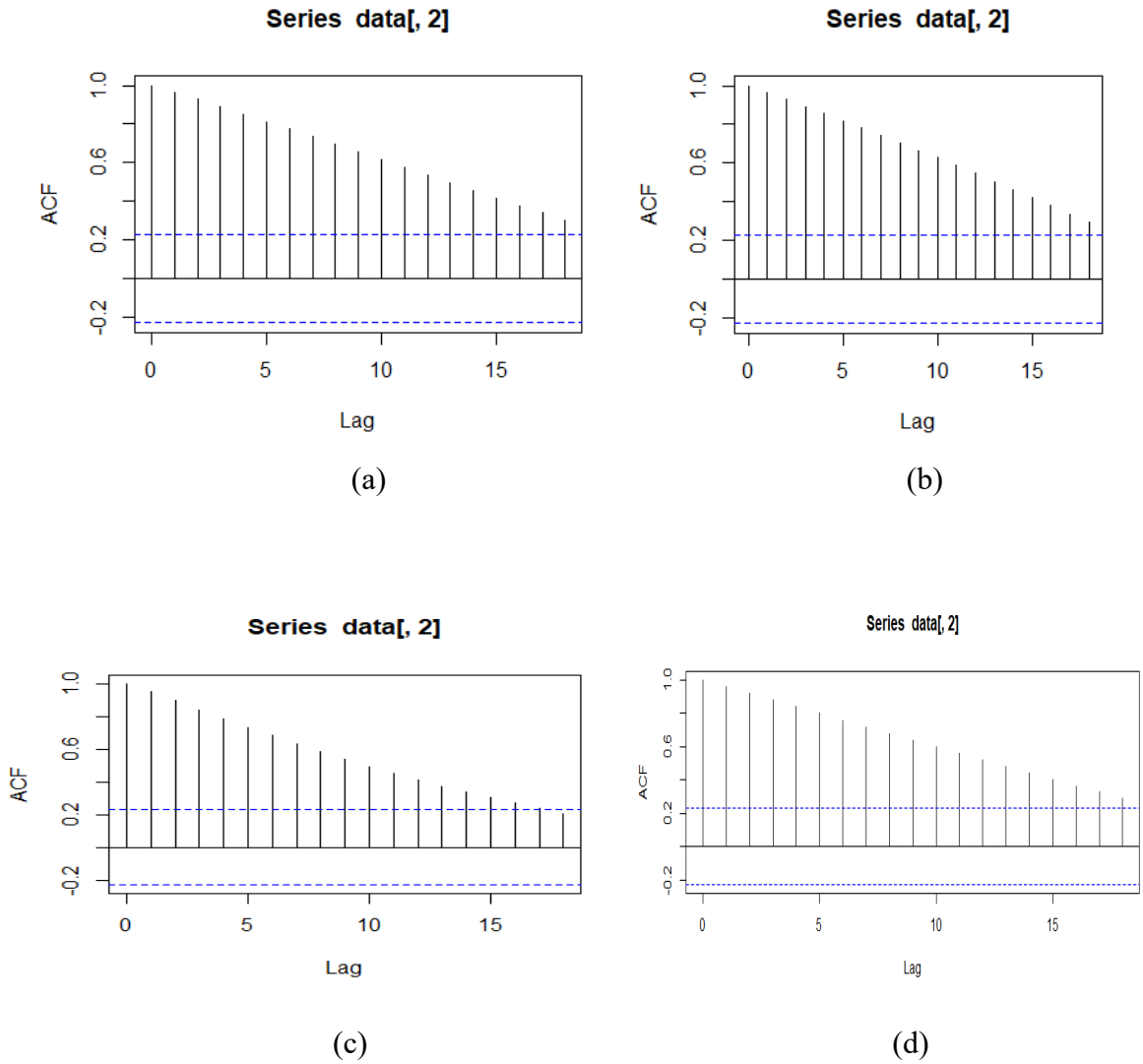

Fig 2.1: The ACF plots for population of (a) India (b) China (c) Russia and (d) United States

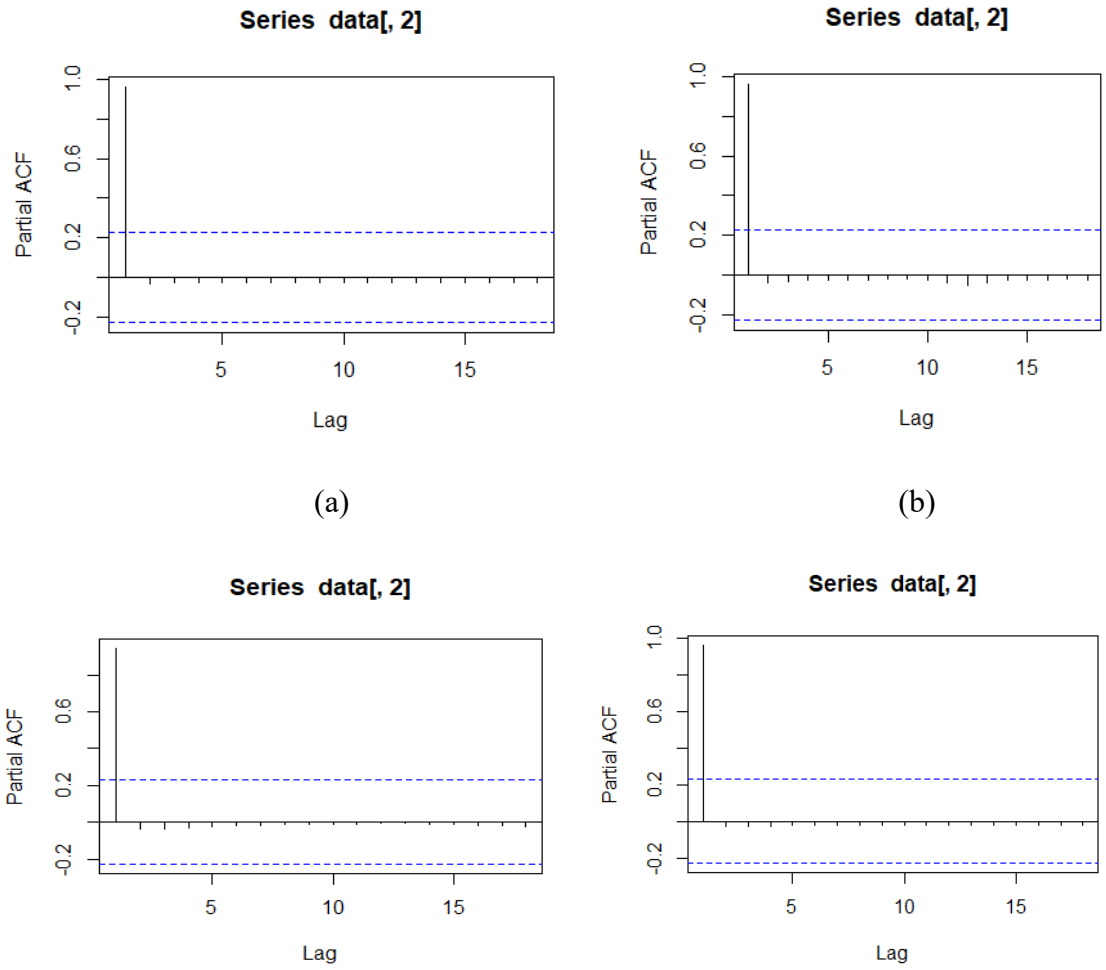

Fig 2.2: The PACF plots for population of (a) India (b) China (c) Russia and (d) United States

2.4 Identification of the Best Model

Identification of the best model has been done using RMSE values. The model with least error is the best model and gives the most accurate prediction. Comparison of different models using RMSE values for different countries have been tabulated in Table 2.2.

	AR MODEL	MA MODEL	ARMA MODEL	ARIMA MODEL
INDIA	0.153180	3.078269	0.153180	**0.001650**
CHINA	0.028514	1.375058	0.021955	**0.019267**
RUSSIA	0.000814	0.130126	0.000814	**0.000711**
UNITED STATES	0.001561	0.531007	0.001561	**0.001212**

Table 2.2: RMSE values for various models for the four countries

From Table 2.2 it is clear that ARIMA Model gives the least error value for all the four countries. Therefore, we can conclude that ARIMA is the best model for predicting the future population of these four countries.

2.5 Fitting of the ARIMA model on the training tests

The training data are utilized to fit the models using the statistical software for R. The parameters of the ARIMA model that gives the best fit, identified for the various countries, are given in Table 2.3.

Country	p	d	q
India	1	2	0
China	2	2	1
Russia	3	2	0
United States	2	2	0

Table 2.3: Parameters of the ARIMA model

Figure (2.3) presents the fitted ARIMA models for the various countries.

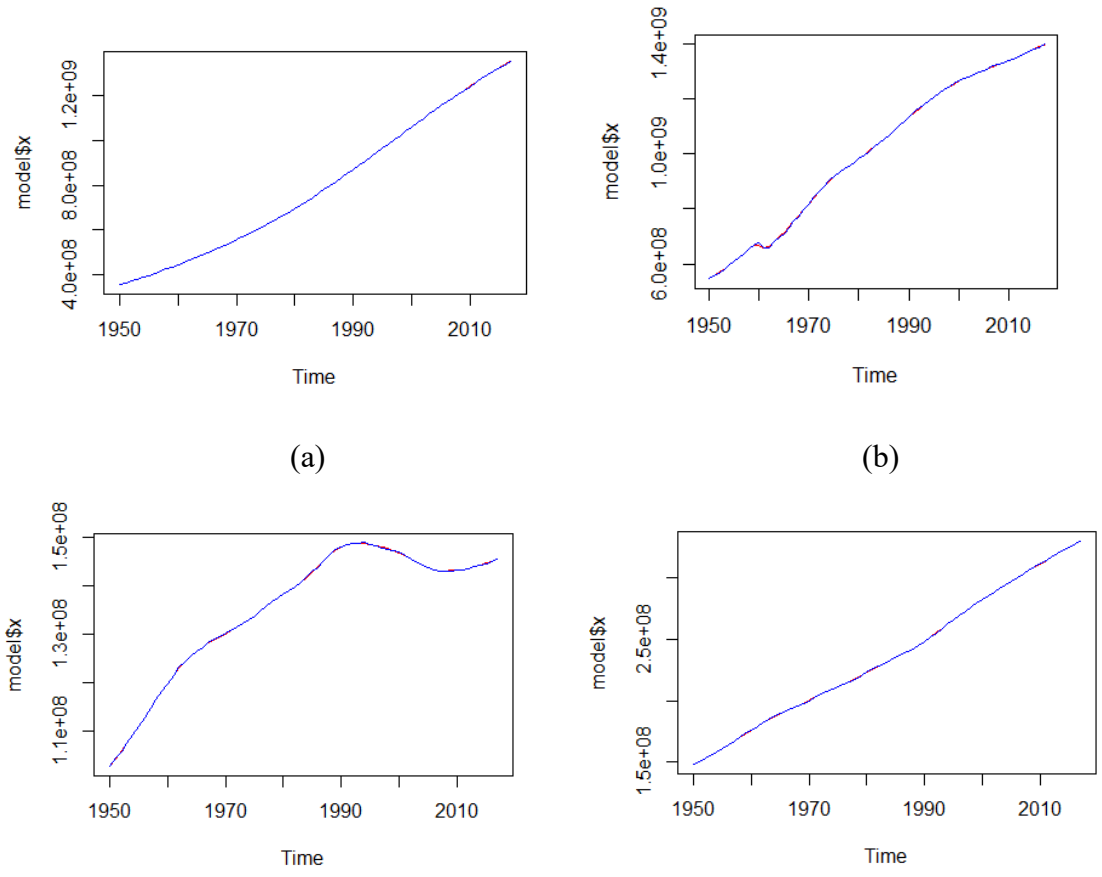

Fig. 2.3: Fitting of ARIMA model for the training sets of (a) India (b) China (c) Russia and (d) United States

2.6 Forecasting of population from 2018 to 2022 on the testing sets using the fitted ARIMA models

Figure 2.4 presents the predicted population of the four countries, i.e., India, China, Russia and United States respectively, from the year 2018 to 2022, along with the original population of these countries. Here, the red line represents the original population and the blue line represents the predicted population. As the model of best fit has been used, the lines are extremely close to each other.

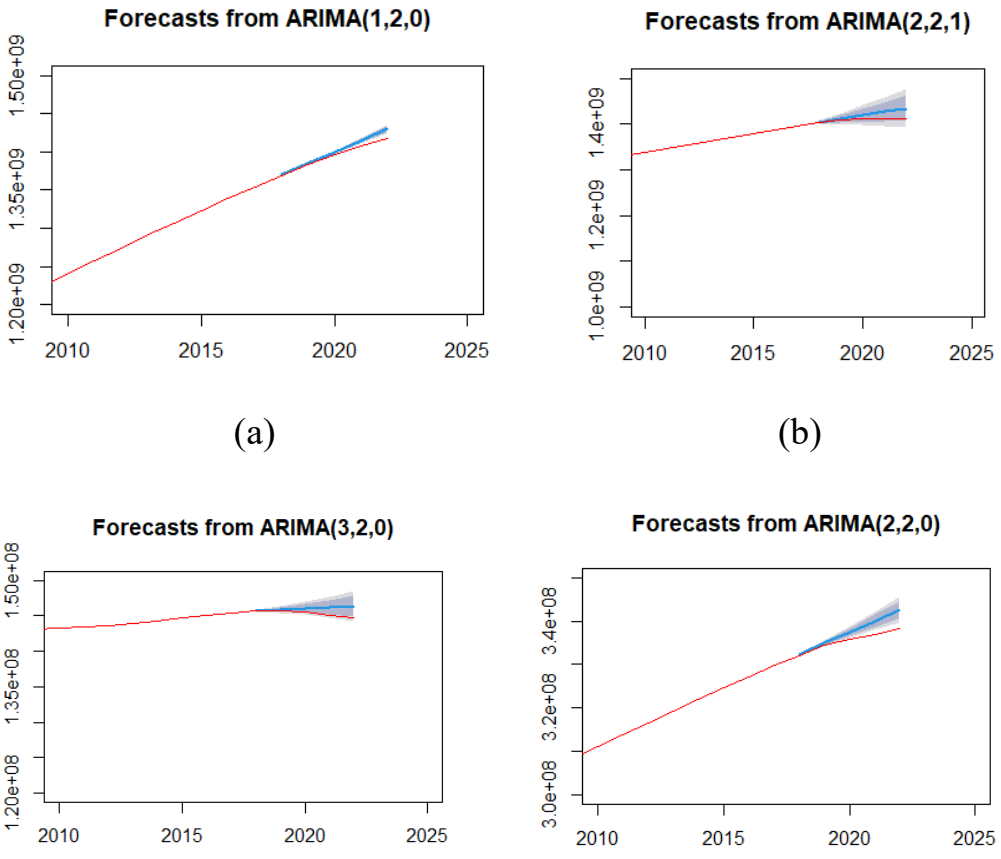

Fig. 2.4: Original and predicted values on the test sets for (a) India (b) China (c) Russia and (d) United States

2.7 Forecasting of Populations in 2025

As ARIMA has been tested to be the best model, it can be used to make short-term predictions of the population. The populations of the four countries in the year 2025 and the percentages of change have been tabulated in Table 2.4.

	PREDICTED VALUES FOR 2025	**PERCENTAGE OF CHANGE***
INDIA	1437919451	1.4639
CHINA	1408767553	-0.2113
RUSSIA	144194015	-0.3588
UNITED STATES	343872192	1.6501
*** Negative values indicate % decrease**		

Table 2.4: Predicted populations of the countries in 2025 and the percentages of change

The results in Table 2.4 indicate the following:

i. For India and United States, there is a population increase, whereas for China and Russia, the population is predicted to decrease.
ii. India is predicted to have the largest population in the year 2025 with a population increase of 1.46%.
iii. Russia is predicted to have the least population with a population decrease of 0.36%.
iv. United States is predicted to have the highest population growth, i.e., 1.65%.

CHAPTER 3

RESULTS AND CONCLUSIONS

3 RESULTS AND CONCLUSIONS

3.1 Conclusions

The present study helped to predict the highest populated country in the year 2025 and the country with the largest population growth. Identification of such trends in population change may help the respective governments to formulate policies and manage resources accordingly.

3.2 Demographic factors

The demographic factors responsible for the decline in population of China and Russia may be explored. A few of them are listed below:

1. **ONE-CHILD POLICY**: In China, One-child policy was implemented in 1979 limiting most urban couples to only one child. Subsequently, it was relaxed to two-child policy and later to three-child policy in 2021.
2. **ECONOMIC FACTORS**: Economic pressures, such as the high cost of education, healthcare, and housing, also the pursuit of career opportunities and economic stability.
3. **AGING POPULATION**: It is due to increased life expectancy and decreased birth rates. Also, it decreases the proportion of working-age individuals.
4. **LOW BIRTH RATE**: Russia has experienced consistently low birth rates over the past few decades.
5. **HIGH MORTALITY RATE**: Alcoholism, smoking, poor healthcare infrastructure in rural areas, and the prevalence of diseases like HIV/AIDS and tuberculosis results in increasing the mortality rate among working-age individuals.
6. **EMIGRATION**: In Russia, lack of economic opportunities and political stability leads to people, especially among skilled workers and young professionals.

4 BIBLIOGRAPHY

[1] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

[2] Dai, J., & Chen, S. (2019). The application of ARIMA model in forecasting population data. In Journal of Physics: Conference Series 1324. IOP Publishing.

[3] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[4] Shumway, R., & Stoffer, D. (2019). Time series: a data analysis approach using R. CRC Press.