ESTIMATION OF FUTURE POPULATION STATUS USING GLOBAL BIRTH RATE ANALYSIS

BY

OSHO EMMANUEL

ABSTRACT

This study aims to analyze global birth rates and their potential implications for future population status. We apply a mix of Multiple Linear Regression Time Series and Exploratory Analysis using a complete dataset that includes variables such as population, birth rate, death rate, fertility rate, population growth rate, and youth dependence ratio. The Exploratory Analysis helps us understand the data better by spotting patterns, trends, and linkages in the data, while Multiple Linear Regression and Time Series enables us to model and predict future population changes based on historical data. The findings of this comprehensive analysis will lead to a more sophisticated understanding of worldwide birth rates and their implications for future population expansion. The findings will help policymakers, researchers, and organizations involved in population planning and policy formation make educated decisions and develop solutions to address population-related concerns.

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Background to the study

The birth rate has played a major part in the significant shifts in the world population that have occurred throughout history. Birth rates are a key factor in determining the future direction of societies, making the study of population dynamics an important component of demographic science (Bongaarts, 2009). Understanding global birth rates and their implications for future population status is crucial for population planners, researchers, and organizations involved in policy development. According to Wikipedia contributors (2023) the global birth rate is defined as the total number of live human births per 1,000 population for a given period divided by the length of the period in years. Using a comprehensive with elements like population, birth rate, death rate, and fertility rate, population growth rate, and youth dependence ratio, this study intends to assess global birth rates and their possible effects on future population expansion.

The twenty-first century has witnessed extraordinary fluctuations in the world population, with the population now exceeding 7 billion and still continuing to rise. This rapid growth, combined with shifting birth rates, poses both challenges and opportunities for communities all around the world. As we look ahead, it is becoming increasingly necessary to examine global birth rates and comprehend their possible effects on population growth. We can foresee and handle any challenges and implications by obtaining insights into these patterns.

The world's population is projected to reach approximately 10.9 billion, with annual growth of less than 0.1% – a steep decline from the current rate (Pew Research Center, 2021), Such a demographic transition has profound effects on a number of societal elements, such as healthcare, education, labor markets, and resource allocation. Examining the world's birth rates is essential to understanding their dynamics and prospective effects on population patterns in the future.

Birth rates have historically been impacted by a wide range of variables, including social, economic, cultural, and environmental determinants (*Factors Affecting Population | Birth Rate, Death Rate, Net Migration*, n.d.). High infant mortality rates, restricted access to contraception, cultural inclinations for bigger families, the need for labor in agricultural economies, and other factors have all contributed to high birth rates in the past. However, birth rates have been steadily declining in many nations as a result of industrialization, urbanization, and improvements in healthcare and education. As a result of decreased birth rates, population dynamics have changed dramatically, and these changes are usually characterized by an aging population and declining youth dependence ratios. These changes have broad ramifications for many sectors of society, including straining healthcare systems as the proportion of the elderly rises and necessitates increased funding for medical and elder care services (Nargund, 2009). On the contrary, regions or localities with high birth rates confront different opportunities and challenges. Rapid population expansion can put a strain on already-scarce resources, raise the rate of poverty, and impede social and economic advancement.

The case study for this project includes the birth rate, death rate, fertility rate, population growth rate, and youth dependence ratio of the following regions: Africa (UN), Asia (UN), Europe (UN), Latin America and the Caribbean (UN), Northern America (UN), Oceania (UN), and the world. In order to provide a thorough picture of global population dynamics and to enable a comparative examination of birth rates and their possible impact on future population trends, these regions were chosen. We can learn more about the parallels, differences, and particular variables affecting birth rates on different continents by looking at these distinct places.

1.2 Historical Background

Birth rates and population dynamics have been researched for centuries and have a substantial historical foundation. Human societies have long recognized the significance of population size and structure in determining how societies grow. Early studies of birth rates and population growth were usually motivated by concerns about labor supply, resource estimation, and the prospect of societal stability or decline.

Plato and Aristotle debated the relationship between population growth and its impact on the social and political fabric of society in ancient Greece, which is one of the earliest recorded discussions on population dynamics (Keeping the Balance: Ancient Greek Philosophical Concerns with Population and Environment on JSTOR, n.d.). However, it wasn't until the 18th century that population studies gained scientific rigor and were recognized as a distinct field of study.

During this time, Thomas Malthus, a British scholar, made important contributions to the study of population. In his influential work, "An Essay on the Principle of Population," published in 1798, Malthus proposed a theory that became known as the Malthusian theory. He claimed that population growth tends to outpace the growth of resources, leading to a struggle for survival and potential societal challenges like poverty, famine, and disease (Dunn, 1998). Malthus' the theory established the groundwork for further research into population dynamics by posing significant issues concerning the connection between birth rates, mortality rates, and resource availability.

Getting to the 19th century, advancement in statistical method and the availability of demographic data facilitated a more comprehensive and quantitative approaches to the study of birth rate. Researchers began to collect and analyze data on births, deaths, and marriages to better understand population trends. The Belgian mathematician and sociologist Adolphe Quetelet's creation of life tables, which allowed for the analysis of mortality rates and the calculation of life expectancies, had a significant impact on the study of birth rates and population projections during this time (Eknoyan, 2007).

With the establishment of the demographic transition theory in the early twentieth century, there was a dramatic shift in understanding of birth rates and their relevance to societal growth. According to this theory, which was put forth by American demographer Warren Thompson, societies move from having high birth and death rates to having low birth and death rates as they experience economic and social development (*Demographic Transition*, n.d.). According to the demographic transition theory, advances in living conditions, healthcare, education, and access to contraception lead to lower birth rates.

The advent of the demographic transition theory was a key milestone in the study of birth rates, shedding light on the complicated interplay of demographic, social, economic, and cultural aspects. This view emphasized that birth rates are substantially influenced by socioeconomic conditions and individual decisions within a society, rather than only biological considerations. The hypothesis allows academics to acquire insights into historical trends of birth rates seen across various countries and regions by providing a complete framework. The demographic transition theory, with its holistic approach, helped to a fuller understanding of the complex mechanisms that determine birth rates and cleared the path for additional research in this field.

The concept of the demographic dividend, developed by David Bloom, David Canning, and Jaypee Sevilla in 2003, has refocused attention on the potential economic benefits that can result from declining birth rates and a change in the age structure of the population. With the world's population now exceeding 7 billion, concerns about high birth rates have given way to issues related to declining birth rates in many countries (Bloom et al., 2003).

This framework emphasizes the economic payoff that can occur when there are relatively large numbers of people of working age. This can happen when substantial declines in fertility lead to a lower share of dependent young children in the population. This can be a boon for economic growth, as there are more workers to support a smaller number of dependents. However, the demographic dividend is not automatic. It requires countries to invest in their human capital by providing quality education and healthcare to their citizens. It also requires countries to create an environment that is conducive to economic growth, such as by investing in infrastructure and reducing corruption.

Countries that can successfully harness the demographic dividend can experience rapid economic growth and improved living standards for their citizens. However, countries that fail to invest in their human capital and create an enabling environment may miss out on this opportunity.

1.3 Aim And Objective Of The Study

The primary objective of this research is to utilize robust analytical methods, such as exploratory analysis, multiple regression, and time series analysis, to examine the birth rate on a global scale and examine its connections with important demographic variables like death rate, fertility rate, population growth rate, and youth dependence ratio. Additionally, the study aims to make predictions for future birth rates based on the identified relationships. Through the application of these advanced statistical techniques, the research strives to improve our comprehension of the intricate dynamics involved in global birth rates and offer valuable insights for population forecasting.

objectives of this study are as follows:

1. Analyze global birth rates and their implications for future population status.
2. Apply Multiple Linear Regression Time Series and Exploratory Analysis to understand the patterns and trends in birth rates.

3. Model and predict future population changes based on historical data using Multiple Linear Regression and Time Series.
4. Provide a comprehensive understanding of worldwide birth rates and their implications for future population expansion.
5. Assist policymakers, researchers, and organizations involved in population planning and policy formation in making informed decisions and developing solutions to address population-related concerns.

1.4 Source Of Data

This study relies on secondary data from Our World in Data to provide a detailed analysis of worldwide birth rates and their implications. The dataset contains data on birth rates, population size, mortality rates, fertility rates, population growth rates, and youth dependence ratios for various areas as well as the entire world. The following data sources were used in this study.

Our World in data: Our World in Data is a reliable online resource that offers extensive worldwide statistics on a variety of socioeconomic and demographic indices. The primary source for examining birth rates across various regions, including Africa (UN), Asia (UN), Europe (UN), Latin America and the Caribbean (UN), Northern America (UN), Oceania (UN), and the entire planet, is the dataset available from Our planet in Data.

Our World in statistics meticulously compiles its statistics from a variety of sources, including international organizations, national statistical organizations, research papers, and official government documents. To guarantee accuracy and dependability, everything goes through stringent quality checks and is updated frequently.

The data can be accessed by visiting the Our World in Data website and searching for "birth rates." The data may also be found by searching the World Bank's World Development Indicators website for "Population, total."

1.5 Limitation Of Study

- The data is from a variety of sources, which may have different methodologies and data
- procedures. This can lead to inconsistencies in the data.
- The study only looks at birth rates at the national level. This means that it does not take into account differences in birth rates within countries.
- The study does not look at the factors that influence birth rates, such as economic development, education, and access to healthcare.
- The study is based on historical data. This means that it may not be accurate for current trends in birth rates.
- The study does not take into account the age structure of the population. This means that it does not account for the fact that some countries have a larger proportion of young people than others. Young people are more likely to have children, so this can skew the results of the study.

CHAPTER 2

LITERATUE REVIEW AND METHODOLOGYS

2.1 Introduction

The chapter focuses on relevant research materials for achieving the objectives of this study and also This chapter outlines the methodology used in this study to analyze global birth rates and their implications for future population status. The study employs a mix of Multiple Linear Regression, Time Series, and Exploratory Analysis. The methodology is divided into four sections main: collection data preprocessing, exploratory data analysis, and multiple linear regression and Time series analysis.

2.2 Backgroud Information On Global Birth Rate

The average number of births per woman throughout the world is referred to as the global birth rate. It is an essential demographic marker that sheds light on worldwide trends in population growth and fertility. Cultural norms, socioeconomic situations, access to healthcare, education, and governmental regulations are just a few of the variables that have an impact on the world's birth rate. For decision-makers, scholars, and organizations concerned in population planning, resource distribution, and sustainable development, understanding worldwide birth rates is crucial. It is possible to spot patterns, evaluate the effectiveness of initiatives, and make well-informed judgments on population dynamics and associated policies by tracking changes in birth rates throughout the world over time.

2.3 Research On Global Birth Rate

The study of population dynamics and their tremendous influence on our planet has ignited the interest of many academics. Researchers have meticulously researched the patterns, effects, and projections around present birth rates and future projections. We hope to understand how these influences shape our global landscape via broad study and various views.

"Changing Values and Falling Birth Rates" by Samuel H. Preston, published in Population and Development Review in 1986, is one specific journal article that adds to this discussion.

Samuel H. Preston's "Changing Values and Falling Birth Rates" summary:

A special supplement of Population and Development Review released Preston's work, "Below-Replacement Fertility in Industrial Societies: Causes, Consequences, Policies," which looks at the effect of changing values on declining birth rates. The 20-page essay is primarily concerned with the social aspects of affluent nations' below-replacement birth rates.

The paper talks about how industrialization, urbanization, and economic growth have affected family planning techniques and conception views. According to Preston, when civilizations go through these changes, conventional beliefs about family size and fertility tend to change, which lowers birth rates.

Furthermore, Preston examines the role of education and women's empowerment in influencing fertility decisions. He highlights how increased access to education, particularly for women, correlates with declining birth rates due to the postponement of marriage and childbearing, as well as increased participation of women in the labor force.

The article "Do High Birth Rates Hamper Economic Growth?" by Hongbin Li and Junsen Zhang, published in The Review of Economics and Statistics in 2007, explores the relationship between birth rates and economic growth and aims to shed light on whether high birth rates have a negative impact on economic development. Li and Zhang use a rigorous analysis that incorporates various factors, including human capital formation, labor supply, and resource allocation to examine the relationship between birth rates and economic growth

(Li & Zhang, 2007).

The analysis by the authors casts doubt on the belief that large birth rates automatically hinder economic development. They argue that there are many variables that might affect the complex connection between birth rates and economic growth. Li and Zhang (2007) support their claims with actual data and statistical models, showing that the influence of high birth rates on economic growth depends on a variety of contextual circumstances and governmental actions.

Li and Zhang (2007) contribute to knowledge of how population dynamics might affect economic results by looking at how birth rates affect important economic variables including labor market outcomes, educational investments, and productivity. Their results have ramifications for scholars and politicians who are interested in developing successful plans that strike a balance between population increase and economic development.

In terms of strengths, Li and Zhang's (2007) work adopts strong techniques and comprehensive statistical analysis, which adds to the confidence of their conclusions. They take into account a wide range of elements and variables to provide a thorough evaluation of the link between birth rates and economic growth. Furthermore, their study provides important insights into the possible policy consequences for regulating population increase and supporting long-term economic development.

However, the limitation of their research was that their research focuses on aggregate-level analyses rather than the micro-level factors underlying the link between birth rates and economic development. Furthermore, the paper focuses exclusively on the setting of China, which may restrict the findings' generalizability to other nations or areas Which was also acknowledge by
Li and Zhang (2007).

Lotka's (1907) study investigates the relationship between birth and death rates and the consequences for population dynamics. Birth rates, which indicate the number of births per unit of population during a certain time period, and death rates, which represent the number of deaths per unit of population over the same time period, are the fundamental concepts under consideration. Lotka examines these rates to better understand their interactions and the consequences for population growth or stability.

In terms of methodology, Lotka's (1907) work is grounded in historical observations and empirical data analysis. Although the article does not explicitly mention specific theories or models, it provides a comprehensive examination of historical birth and death rates to uncover any patterns or correlations. The study's empirical approach enhances its credibility and allows for a better understanding of population dynamics during the time period in question

Even though Lotka's (1907) study does not adopt an innovative approach, it remains a fundamental paper. It helps to understand population dynamics by providing a historical perspective and insights into the dynamics of birth and death rates. Lotka's research established the framework for later studies in the subject, influencing the direction of research on demographic trends and their societal ramifications.

Lotka's (1907) study is notable for its historical relevance as well as its contributions to the subject of population dynamics. The study sheds light on the historical backdrop of population growth and stability by studying birth and death rates. However, it is critical to recognize the limits of historical research, such as potential data gaps, restricted data availability, and dependence on historical documents.

In terms of its relation to other literature, Lotka's (1907) article serves as a pivotal publication in the field of population dynamics. While subsequent sstudies havebeen built upon Lotka's findings or challenged certain aspects.

Studies such as The Impact of Population Ageing on Economic Growth: A Comparative Analysis of OECD Countries" by Andrea Baldacci, Michele Ciccarelli, and Andrea Lanza (2012), "The Impact of Population Growth on Environmental Sustainability" by Sangwon Suh and David J. Teece (2013),

"The Impact of Population Growth on Social Inequality" by David E. Bloom, David Canning, and Jaypee Sevilla (2014),

"The Impact of Population Growth on Political Stability" by Michael J. Tierney (2015),

"The Impact of Population Growth on Innovation" by Bart Verspagen and Arnoud W. Boonstra (2016).

His work remains a crucial reference point for understanding the historical development of population dynamics research.

2.4 Data Collection

The dataset used in this primarily sourced from Our World in Data (*Our World in Data*, n.d.), a reputable online platform renowned for its comprehensive global statistics on various socioeconomic and demographic variables. This trustworthy source provided valuable data on population, birth rate, death rate, fertility rate, population growth rate, and youth dependency ratio, all of which were included in the dataset.

To further enrich the dataset, additional data points were obtained exclusively from Our World in Data. These supplementary data points, obtained from the same reliable source, were meticulously incorporated into the existing dataset. This process aimed to enhance the dataset's comprehensiveness and enable a more comprehensive analysis of the research objectives.

Any possible problems with data compatibility or inconsistencies were minimized by carefully integrating the extra data from the same source. The combined dataset that results keeps the dependability and trustworthiness of Our World in Data (*Our World in Data*, n.d.), guaranteeing the correctness and integrity of the information gathered.

2.5 Data Preprocessing

Data preprocessing, which involves cleaning the data and converting it into a format that can be easily analyzed, is an important step in any data analysis process (Kelleher et al., 2015). In this study, the data preprocessing included the following procedures:

Checking for missing values: Any missing values in the dataset would be identified and handled appropriately. In some cases, missing values would be filled using appropriate statistical methods, such as mean or median imputation or using Forward or Backward fill. In other cases, rows with missing values would be dropped from the dataset.

Checking for outliers: Outliers, or data points that deviate considerably from the rest of the dataset, would be detected and dealt with. Outliers can skew analytical results, thus they must be identified and dealt with carefully.

Normalization: To verify that all variables were on the same scale, the data would be normalized. This is critical for multiple linear regression analysis because it assures that all variables have an equal opportunity to influence the outcome.

Handling unbalanced data: If the dataset has a class imbalance, strategies like oversampling the minority class or under sampling the majority class would be used to resolve the issue and provide a balanced representation of the classes.

Feature engineering: To improve the model's predictive capacity, new features would be constructed from existing ones. Creating interaction terms, polynomial characteristics, and generating new variables based on domain knowledge were all part of this.

Skewed data distributions would be handled by using transformations such as logarithmic or power transformations to normalize the distribution and improve the analysis's effectiveness.

Managing multicollinearity: To handle multicollinearity, variance inflation factor (VIF) analysis or principal component analysis (PCA) were used to discover and eliminate strongly linked variables.

2.6 Exploratory Data Analysis

Exploratory Data Analysis (EDA), which is essentially a way to summarize and visualize the key characteristics of a dataset before making assumptions or building statistical models, is a crucial step in the data analysis process that allows for a better understanding of the patterns, trends, and relationships in the data (Beyer, 1981). In this study, EDA was used to gain insights into the data and direct the subsequent analysis.

The Exploratory Data Analysis (EDA) process in this study involved the following steps:

2.6.1 Descriptive Statistics

Descriptive statistics are an important part of EDA since they allow you to summarize and characterize the major properties of the dataset. It assists you in understanding the fundamental aspects of the data before delving deeper into the research. Consider the following essential descriptive statistics:

- Central tendency measures: mean, median, and mode,
- Dispersion measures include standard deviation, range, and interquartile range,
- Histograms, bar plots, or pie charts for frequency distribution.

2.6.2 Data Visualization

Data visualization is important in EDA because it helps bring data to life and allows us to visually find patterns and trends. Depending on the type of data and the insights desired, many visualization approaches will be utilized.

- Scatter plots: used to show the connection between two variables,
- Box plots: to identify the spread and outliers in a dataset,
- Line graphs: to spot trends and changes over time.
- Heatmaps: to display correlations between different variables.

2.6.3 Correlation Analysis

Correlation analysis helps us to determine the degree of relationship between two or more variables. It aids in the identification of linear correlations and can help forecast one variable based on another. Correlation analysis that would be used include:

- Pearson correlation coefficient quantifies the degree and direction of a linear link between two variables,
- Scatter plots: graphically display the relationship between variables.

2.6.4 Trend Analysis

Trend analysis would be conducted to identify patterns or trends in the data over time. This is particularly important for time series data, as it can reveal patterns that are not immediately apparent from the raw data and it helps in understanding how the data changes and if there are any recurring patterns. Techniques used in trend analysis are:

- Line graphs: visualize how variables change over time,
- Moving averages: smooth data to identify long-term trends by reducing noise.

2.6.5 Outlier Detection

Outliers, or observations that deviate significantly from the other observations, can distort the results of the analysis. Box plots and scatter plots were used to visually identify outliers. Additionally, statistical methods used for detect outliers include:

- Box plots: visually identify observations outside the whiskers,
- Z-score: calculate the number of standard deviations an observation is from the mean,

- Tukey's fences: define boundaries for identifying outliers based on the interquartile range.

Through EDA, a comprehensive understanding of the dataset would be achieved, which guided the subsequent steps of the analysis. It will help in identifying potential issues such as outliers and multicollinearity and provided insights into the relationships between variables. These insights would be crucial in informing the choice of modeling techniques and the interpretation of the results

2.7.1 Multiple Linear Regression

Using the values of two or more independent variables, multiple linear regression is a statistical method for predicting the outcome of a dependent variable (Krzywinski & Altman, 2015). In this study, Multiple Linear Regression was used to model and predict future population-based changes on historical data.

A multiple linear regression model has the following general form:

$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta nXn + \varepsilon$

where:

- Y is the dependent variable (in this case, the birth rate)
- β0 is the y-intercept
- β1, β2, ..., βn are the coefficients of the independent variables
- X1, X2, ..., Xn are the independent variables (in this case, the population, death rate, fertility rate, population growth rate, and youth dependence ratio)

- ε is the error term

The coefficients (B1, B2, …, Bn) would be calculated using the least squares approach, which minimizes the sum of squared residuals (the discrepancies between observed and predicted values of the dependent variable).

These coefficients express the change in the birth rate caused by a one-unit change in each independent variable. Finally, denotes the error factor, which accounts for the variation in the birth rate that cannot be explained by the independent variables.

2.7.2 Model Evaluation

The performance of the multiple linear regression model was evaluated using various statistical measures, such as the R-squared value, which indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. A high R-squared value indicates a good fit of the model to the data.

In addition, the residuals (the differences between the observed and predicted values of the dependent variable) were analyzed to check for any patterns that might suggest a problem with the model.

The findings of the multiple linear regression analysis would shed light on the association between birth rates and the independent factors. The computed coefficients would reveal the size and direction of each independent variable's influence on the birth rate. A positive coefficient for population, for example, indicates that a rise in population is connected with higher birth rates, whereas a negative coefficient for death rate indicates that an increase in death rates is associated with lower birth rates.

2.8 Time Series Analysis

A statistical technique called time series analysis is used to examine and project data points gathered over a period of time (Liu et al., 2021). Time series analysis is used in this study to analyse historical birth rate data and create predictions about future birth rates based on observed trends and patterns.

The following stages are included in the time series analysis in this study:

2.8.1 Data Preparation

The birth rate would be dependent variable and the relevant time periods would be the independent variables. The dataset would then be organized in chronological order. The data was processed properly after being reviewed for any missing values or discrepancies.

2.8.2 Stationarity Testing

A crucial presumption in time series analysis is stationarity, which states that the data's statistical characteristics are constant across time. Tests like the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test would be performed to make sure stationarity. In order to attain stationary, if the data were determined to be non-stationary, procedures like differencing or logarithmic transformations would be used.

2.8.3 Trend And Seasonality Analysis

Analysis: While seasonality analysis finds recurrent patterns within a certain time period, trend analysis helps uncover long-term patterns or trends in the data. Moving averages, exponential smoothing, and decomposition are some examples of techniques that would be utilized to analyze and visualize patterns and seasonality in the data.

2.8.4 Model Selection

Several time series models exist time series model such as Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Exponential Smoothing State Space Model (ETS), and Vector Autoregression (VAR) etc. But the model that would be used in this study is the Autoregressive Integrated Moving Average (ARIMA). Moving average (MA), differencing (I), and autoregressive (AR) are the three primary components of ARIMA models. Each element helps to capture various facets of the time series data.

1. autoregressive (ar) component

This component models the relationship between the variable and its lagged values. It captures the impact of previous observations on the time series' present value. The parameter p, which specifies the number of lagged data incorporated in the model, constitutes the AR component of ARIMA.

The autoregressive equation for an AR(p) component is given by:

$Y(t) = c + \varphi_1 Y(t-1) + \varphi_2 Y(t-2) + ... + \varphi_p Y(t-p) + \varepsilon(t)$,

where:

Y(t) represents the value of the time series at time t, c is a constant term,

$\varphi_1, \varphi_2, ..., \varphi_p$ are the autoregressive coefficients for the lagged values,

$\varepsilon(t)$ is the error term or residual at time t

2. moving average (ma) component

The moving average component simulates the reliance on previous forecast mistakes. It captures the effect of earlier error terms on the time series' current value. The parameter q, which reflects the amount of lagged forecast errors contained in the model, represents the MA component of ARIMA.

The moving average equation for an MA(q) component is given by:

$$Y(t) = c + \theta_1\varepsilon(t-1) + \theta_2\varepsilon(t-2) + ... + \theta_p\varepsilon(t-q) + \varepsilon(t),$$

where:

$\theta_1, \theta_2, ..., \theta_p$ are the moving average coefficients for the lagged forecast errors,

$\varepsilon(t)$ represents the error term at time t.

3. differencing (i) component:

ARIMA's differencing component is used to tackle non-stationarity in time series data. Non-stationarity occurs when the statistical features of data, such as mean or variance, alter with time. Differencing is the process of transforming data into a stationary series by calculating the difference between successive observations.

The differenced equation for an integrated component is given by:

$$\Delta Y(t) = Y(t) - Y(t-1),$$

where:

$\Delta Y(t)$ represents the differenced series at time t.

differencing is applied repeatedly until the resulting series becomes stationary.

4. Autoregressive Integrated Moving Average Models

ARIMA models, or Autoregressive Integrated Moving Average models, as previously stated, are commonly employed in time series analysis for forecasting and data analysis. When the data does not show apparent seasonal trends, they are a common choice. ARIMA models integrate autoregressive (AR), differencing (I), and moving average (MA) components to describe the time series' underlying dynamics.

ARIMA (p, d, q) is the generic form of an ARIMA model, were

- p specifies the order of the autoregressive component, which reflects the connection between the variable and its lagged values.
- d is the amount of differencing done to the data in order to attain stationarity.
- q is the order of the moving average component, which models the reliance on previous forecast mistakes.

The ARIMA model equation is as follows:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

6. Seasonal Autoregressive Integrated Moving Average Models

SARIMA, or Seasonal Autoregressive Integrated Moving Average models, is an extension of the ARIMA models that account for seasonal trends in time series data. When analyzing data with clear seasonal patterns, SARIMA models are a suitable choice for forecasting and data analysis. SARIMA models combine autoregressive (AR), differencing (I), and moving average (MA) components, along with seasonal AR, seasonal differencing, and seasonal MA components to capture the underlying dynamics of seasonal time series data.

SARIMA (p, d, q) (P, D, Q, s) is the generic form of a Seasonal Autoregressive Integrated Moving Average model, where:

p represents the order of the autoregressive component for the non-seasonal part, capturing the relationship between the variable and its lagged values.

d is the amount of differencing applied to achieve non-seasonal stationarity in the data.

q denotes the order of the moving average component for the non-seasonal part, modeling the dependence on previous forecast errors.

P signifies the order of the seasonal autoregressive component, which accounts for the relationship between the variable and its seasonal lagged values.

D represents the amount of seasonal differencing performed to achieve seasonal stationarity in the data.

Q indicates the order of the seasonal moving average component, capturing the dependence on previous seasonal forecast errors.

s denotes the number of time steps in a single seasonal period (e.g., 12 for monthly data with a yearly seasonality).

The SARIMA model equation is as follows:

$$\hat{y}t = \mu + \Phi 1yt-s + \ldots + \Phi Pyt-ps - \theta 1et-s - \ldots - \theta Qet-Qs + \phi 1yt-1 + \ldots + \phi pyt-p - \theta 1et-1 - \ldots - \theta qet-q,$$

where $\hat{y}t$ represents the forecasted value at time t, yt represents the observed value at time t, et represents the forecast error at time t, μ is the mean, $\Phi 1$ to ΦP are the seasonal autoregressive coefficients, $\phi 1$ to ϕp are the non-seasonal autoregressive coefficients, $\theta 1$ to θQ are the seasonal moving average coefficients, and $\phi 1$ to ϕq are the non-seasonal moving average coefficients.

2.8.5 Model Fitting And Evaluation

Model Fitting and Evaluation: The data was fitted to the specified time series model, and the model's parameters would be calculated. Statistical measurements such as the Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) will be used to assess the model's performance. The model's residuals were also examined for patterns or autocorrelation.

2.8.6 Forecasting

After fitting and testing the ARIMA and SARIMA model using birth rate data, the model is used to estimate future values. The model's estimated parameters are used to make forecasts for future time periods. These projected values are then compared to the actual values to determine the model's accuracy and dependability.

2.9 Limitations

2.9.1 data availability

The study is reliant on the availability and dependability of Our World in Data. If the dataset has constraints or gaps, the accuracy and comprehensiveness of the conclusions may suffer. Furthermore, specific variables or nations may be missing from the dataset, limiting the generalizability of the conclusions.

2.9.2 assumptions and simplifications

The analysis is based on numerous assumptions, including the linearity of relationships, error independence, and the lack of multicollinearity in the multiple linear regression model. These assumptions may not hold true in all circumstances, and any breaches may have an influence on the results' validity.

2.9.3 stationarity assumption

The ARIMA model implies that the time series data is stationary or can be rendered stationary by differencing. Real-world data, on the other hand, frequently demonstrates trends, seasonality, or other types of non-stationarity that the model may not fully represent. This constraint may have an influence on forecast accuracy and outcomes interpretation.

2.9.4 Generalizability

The findings of this study may be particular to the dataset utilized and may not be entirely generalizable to all areas or time periods. Birth rates can be impacted by a variety of socio-cultural, economic, and policy factors that vary among nations and throughout time. As a result, when applying the findings to diverse circumstances, care should be used.

2.9.5 External Validity

Although the study focuses on worldwide birth rates, the results may not completely represent the unique dynamics and intricacies of specific nations or areas. Birth rates can be greatly influenced by cultural, economic, and social variations, and a worldwide study may ignore key details at the local level

CHAPTER 3

DATA PRESENTATION, ANALYSIS AND DISCUSION

3.1 Introduction

This chapter present the analysis of the global birth rate, data collected from Our World in data from 1950 to 2021 which involves the birth rate, death rate, fertility rate, population growth rate, and youth dependence ratio of the following regions: Africa (UN), Asia (UN), Europe (UN), Latin America and the Caribbean (UN), Northern America (UN), Oceania (UN), and the world. The Python Software was used for the Exploratory Analysis, the Multiple Linear Regression and the various tentative Time series ARIMA models developed was fitted to each data and the suitable models were selected based on diagnostics on the residual of each model and other criterial.

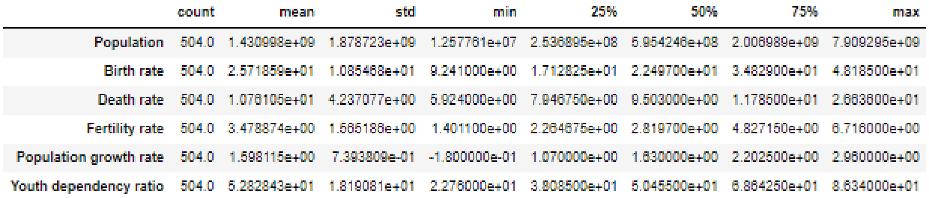
3.2 Data Cleaning And Preprocessing

The raw dataset obtained contained annual data from 1950 to 2021 for the following variables across all continents and the world: Birth rate, Death rate, Fertility rate, Population growth rate, Youth dependency ratio, and Total population.

In addition, the dataset included the categorical variables 'Continent' and 'Year'. Before analysis commenced, the data required cleaning and preprocessing into a suitable format.

The raw data was loaded into a Pandas DataFrame in Python for ease of manipulation and analysis. The info() method was applied on the DataFrame to verify the expected column names and data types matched the imported data. Summary statistics of the DataFrame were computed using the .describe() method to check for any anomalous values or outliers. The quantitative columns had reasonable mean values and ranges, with no abnormal entries. The data was determined to have integrity based on the descriptive overview.

Table 3.1



	count	mean	std	min	25%	50%	75%	max
Population	504.0	1.430998e+09	1.878723e+09	1.257761e+07	2.536895e+08	5.954246e+08	2.006989e+09	7.909295e+09
Birth rate	504.0	2.571859e+01	1.085468e+01	9.241000e+00	1.712825e+01	2.249700e+01	3.482900e+01	4.818500e+01
Death rate	504.0	1.076105e+01	4.237077e+00	5.924000e+00	7.946750e+00	9.503000e+00	1.178500e+01	2.663800e+01
Fertility rate	504.0	3.478874e+00	1.565186e+00	1.401100e+00	2.264675e+00	2.819700e+00	4.827150e+00	6.716000e+00
Population growth rate	504.0	1.598115e+00	7.393809e-01	-1.800000e-01	1.070000e+00	1.630000e+00	2.202500e+00	2.960000e+00
Youth dependency ratio	504.0	5.282843e+01	1.819081e+01	2.276000e+01	3.808500e+01	5.045500e+01	6.864250e+01	8.634000e+01

The 'Year' column contained integer values from 1950 to 2021 representing the time dimension. For temporal analysis, 'Year' needed conversion to datetime format. The Pandas to_datetime() function was used to transform the integers into datetime64[ns] values.

The column names were also cleaned up to be concise and consistent. The long-form names were replaced with simplified versions preserving the key variables. The cleaned column names were:

'Year', 'Continent', 'Population', 'Birth rate', 'Death rate', 'Fertility rate', ' Population growth rate ', and 'Youth dependency ratio'.

With datetime formatted years and simplified column names, the variable types were verified to ensure appropriate formats. 'Year' was datetime64[ns], 'Continent' was object (categorical), and the remaining quantitative columns were float64.

The columns were then rearranged to place 'Year' and 'Continent' first followed by the numerical variables. This logical structure improved analysis workflow. Duplicate entries were checked for using the DataFrame .duplicated() method, but no duplicates were found in the cleaned dataset.

Finally, the processed DataFrame was exported to a CSV file for permanent storage and reusability. The CSV file provided easy data access without repeating the cleaning workflow.

3.3 Exploratory Data Analysis

Exploratory data analysis was conducted to understand distributions, variable relationships, and temporal patterns through extensive visualization and statistical analysis.

3.3.1 Univariate Visual Analysis

To study individual variable distributions, histograms were plotted for each numerical column excluding 'Population'. Figures 3.1 to 3.5 show the histogram plots with normal density curves overlaid.

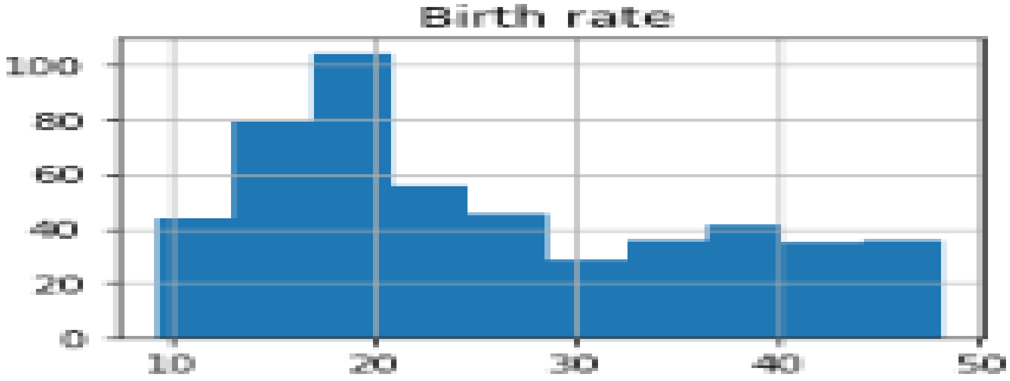


Figure 3.1: Distribution of birth rate

The birth rate distribution in Figure 3.1 demonstrates the downward global trend over time, with an right skewed density peak between 10 to 25 births per 1000 population.

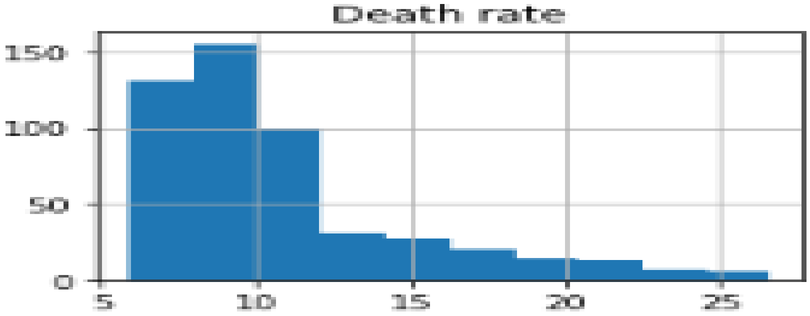


Figure 3.2: Distribution of death rate

The death rate histogram in Figure 3.2 exhibits a left skewed distribution, with a mean under 10 deaths per 1000 population.

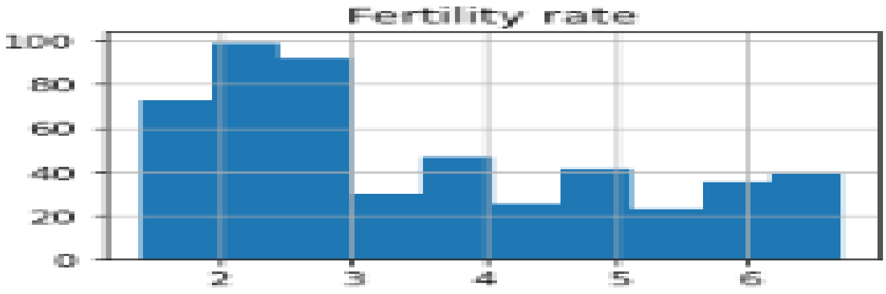


Figure 3.3: Distribution of fertility rate

Figure 3.3 shows the fertility rate distribution shifting left from its peak between 5 to 6 children per woman in the 1950s, indicating declining global fertility over time.

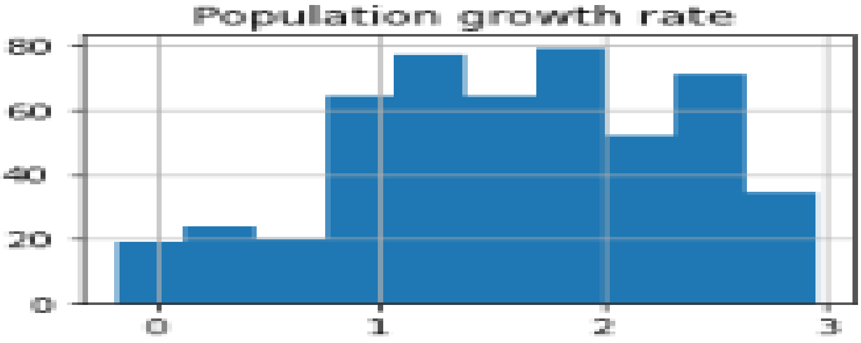


Figure 3.4: Distribution of population growth rate

The population growth rate distribution in Figure 3.4 is relatively normal with a mean around 1.5 to 2.

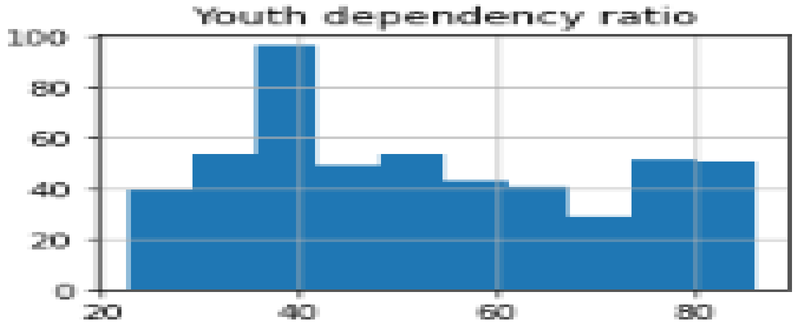


Figure 3.5: Distribution of youth dependency ratio

Lastly, Figure 3.5 displays a roughly normal distribution for the youth dependency ratio, with a mean just over 50% historically.

These histograms provided univariate perspectives into the data distributions. The overlaid density curves confirmed approximate normality of the numeric columns. With this univariate analysis complete, bivariate relationships were explored next.

3.3.2 Correlation Analysis

Correlation analysis was conducted to quantify the strength of linear relationships between variables and identify potential issues like multicollinearity that impact modeling.

First, Figure 3.6 presents the correlation heatmap of all numeric columns excluding 'Population'. Strong positive correlations emerged:

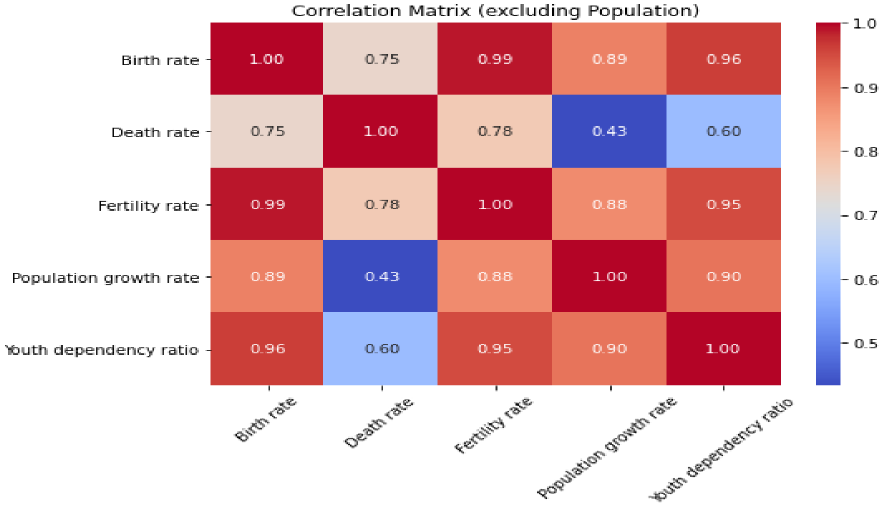


Figure 3.6

Figure 3.6: Correlation heatmap of variables

- Birth_rate has a very high 0.99 correlation with Fertility_rate. This aligns with domain knowledge since higher fertility directly causes increased births.
- Birth_rate and Youth_ratio also correlate strongly at 0.96. Again, this matches demographic theory of more births occurring with greater youth population.
- Growth_rate and Birth_rate exhibit a 0.89 correlation. Increasing births raise the population growth rate.
- Death_rate has a moderately high 0.75 correlation with Birth_rate, which is reasonable given mortality impacts population dynamics.

However, these correlations above 0.75 indicate substantial multicollinearity. This high correlation makes isolating the distinct influence of each predictor difficult in a combined regression model. Their effects are tightly coupled based on the underlying demographic variables. Still, these relationships match expected domain patterns.

3.3.3 Global Trends Visualization

To assess worldwide trends, line plots were created for the key variables over time. Figure 3.7 displays the global birth rate from 1950 to 2021. A steep declining trend is evident, particularly between 1960 and 1990. The birth rate peaked around 35 in 1960 before dropping to under 20 by 2021.

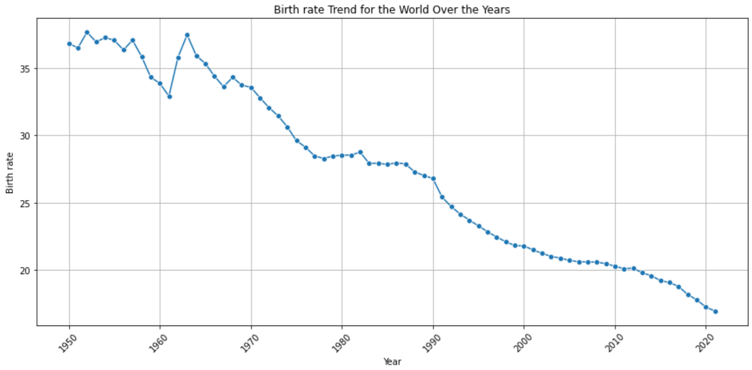


Figure 3.7: Global birth rate from 1950 to 2021

The fertility rate, depicted in Figure 3.8, shows a very similar downward pattern to the global birth rate over the decades. The fertility rate measures the average number of children per woman. Women globally had on average more than 5 children each in 1950, dropping to about 2.5 children by 2021. This massive reduction in fertility rate is a key driver of the declining birth rate observed.

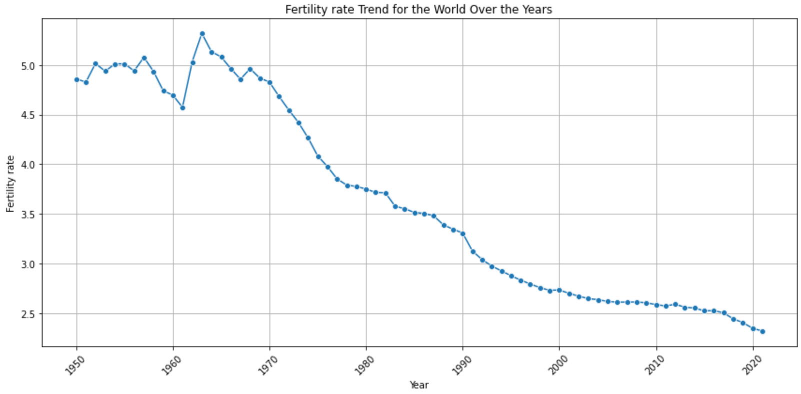


Figure 3.8: Global Fertility Rate from 1950 to 2021

Youths aged 24 years or under as a proportion of the adult population is captured by the youth dependency ratio. As seen in Figure 3.9, the global youth dependency ratio has also decreased over time, most rapidly between 1970 and 1990. A higher youth dependency ratio is associated with a higher birth rate, since it represents large youth cohorts relative to working-age adults.

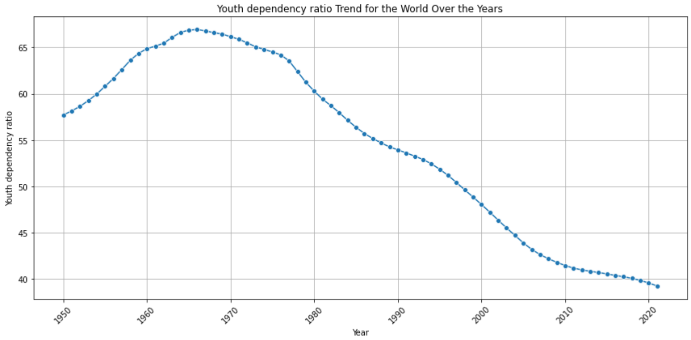


Figure 3.9: Global Youth Dependency Ratio from 1950 to 2021

Despite the reduced birth rate, the total world population has steadily risen over time as depicted in Figure 3.10. But the rate of growth has slowed in recent decades. Population surged from under 2.5 billion in 1950 to over 7.5 billion by 2017.

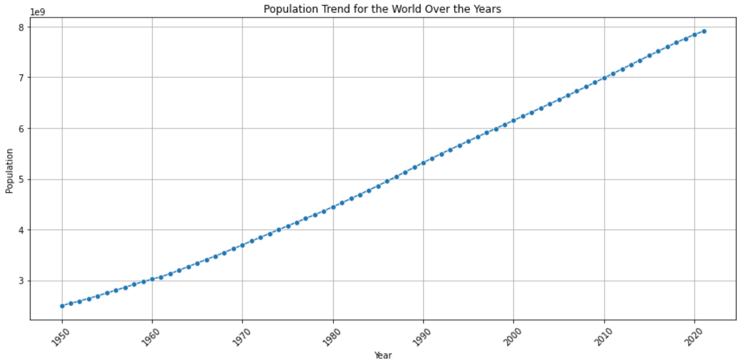


Figure 3.10: Global population growth from 1950 to 2021

Lastly, the global death rate shown in Figure 3.11 declined from around 25 in the 1950s to under 10 by 2021. Medical improvements increasing life expectancy are a key factor in the falling death rate.

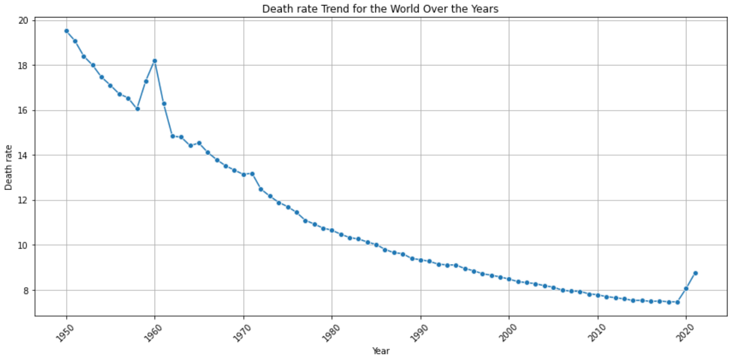


Figure 3.11: Global death rate from 1950 to 2021

3.3.4 Exploring Trends By Continent

To enable comparison between regions, line plots showing time trends were generated for each continent individually, excluding the World aggregate.

Figure 3.12 displays the trends in birth rate by continent from 1950 to 2021. Asia and Africa historically had much higher birth rates peaking around 50 in the 1960s before declining. Africa maintains the highest birth rate reaching near 35 in 2021. Asia's birth rate dropped steeply from over 40 down to 15.

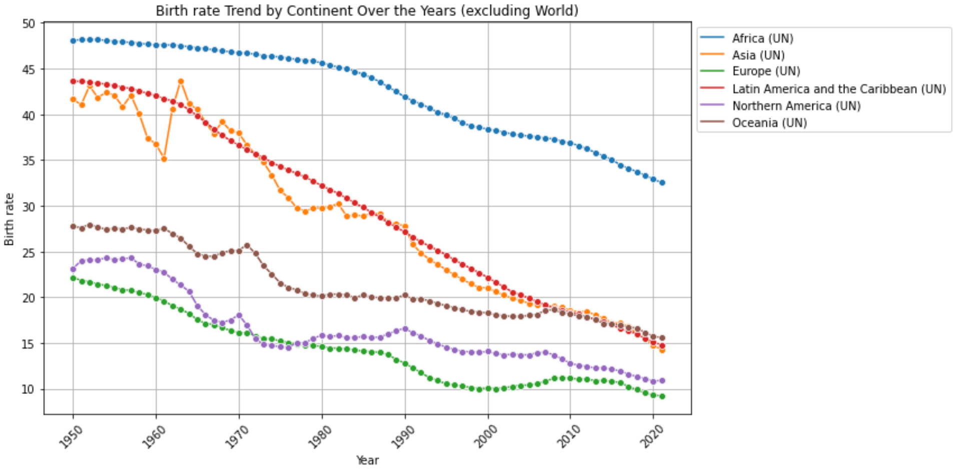


Figure 3.12: Birth rate by continent from 1950-2021

Interesting patterns emerge when comparing fertility rate trends in Figure 3.13. Africa and Asia had substantially higher fertility historically, indicating women had many more children on average. Asia's fertility rate has reduced over time to reach close to the global average. Africa retains the highest fertility rate among continents.

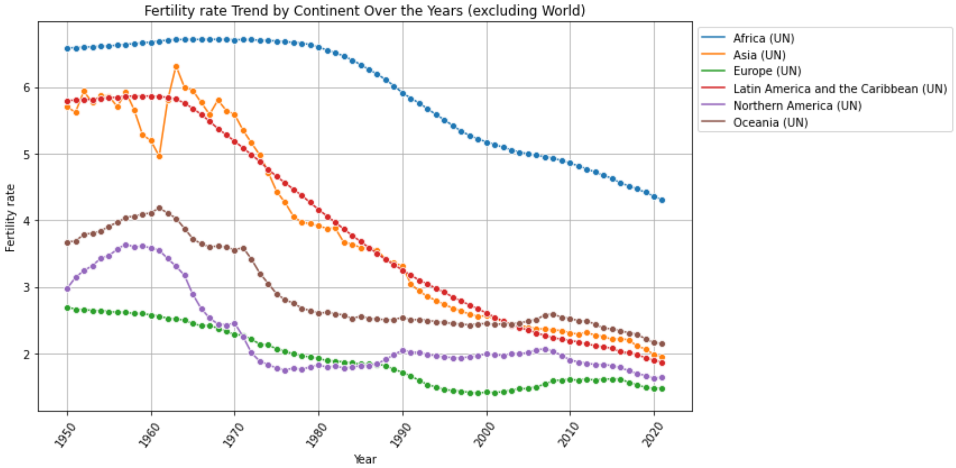


Figure 3.13: Fertility rate by continent from 1950-2021

Figure 3.14 shows the youth dependency ratio by continent from 1950 to 2021. Africa maintains the highest proportion of youth population while Asia's youth dependency has fallen quickly to Northern America and Europe's range. This aligns with Asia's declining birth and fertility rates.

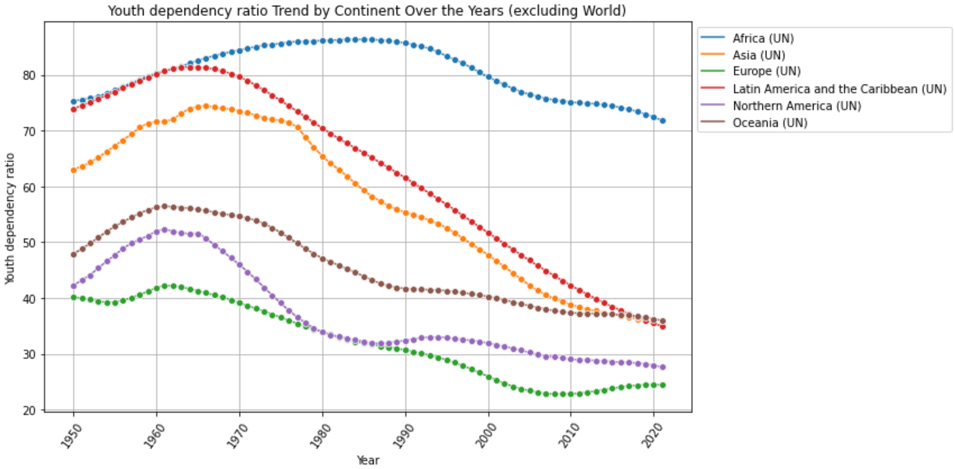


Figure 3.14: Youth dependency ratio by continent from 1950-2021

The death rate comparison by continent is provided in Figure 3.15. In the mid-20th century, Africa had substantially higher death rates reaching around 28. Asia also peaked around 25 in the 1960s. Death rates declined across all continents over the decades.

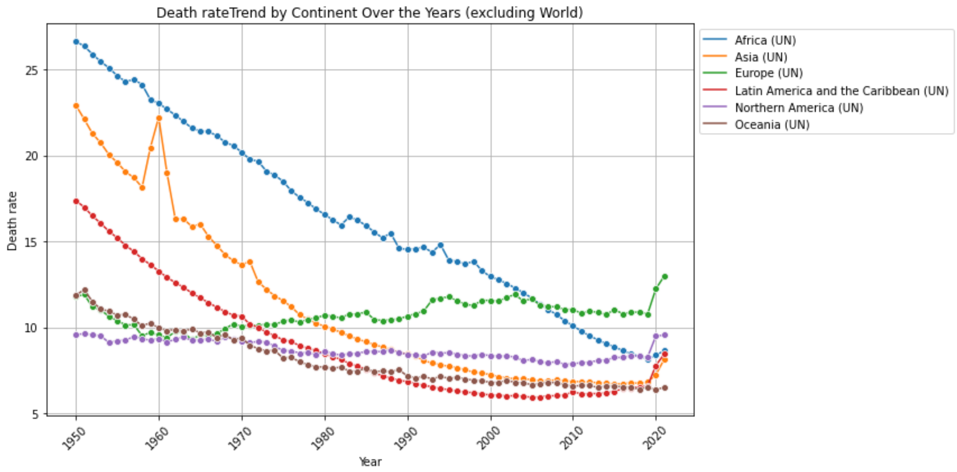


Figure 3.15: Death rate by continent from 1950-2021\

Lastly, Figure 3.16 graphs the total population growth by continent since 1950. Asia experienced massive population growth, increasing from under 1.5 billion to over 4.5 billion people. Africa's population nearly tripled over this period.

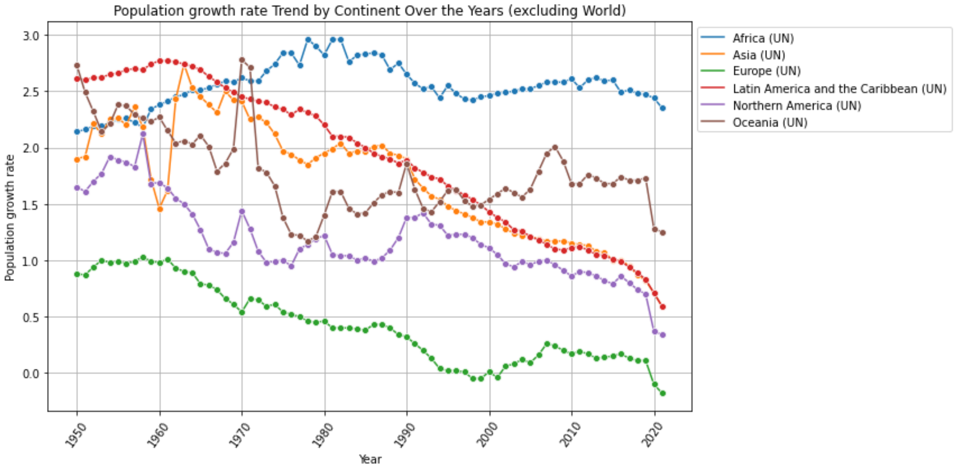


Figure 3.16: Population by continent from 1950-2021

In Conclusion analyzing trends by continent revealed substantial differences between regions including:

- Birth rate declining across all continents, but most steeply in Asia
- Fertility rate historically highest in Africa and Asia, but Asia decreased close to global average
- Youth dependency remains highest in Africa, while Asia's youth ratio has fallen quickly
- Asia's population grew tremendously, Africa's population nearly tripled

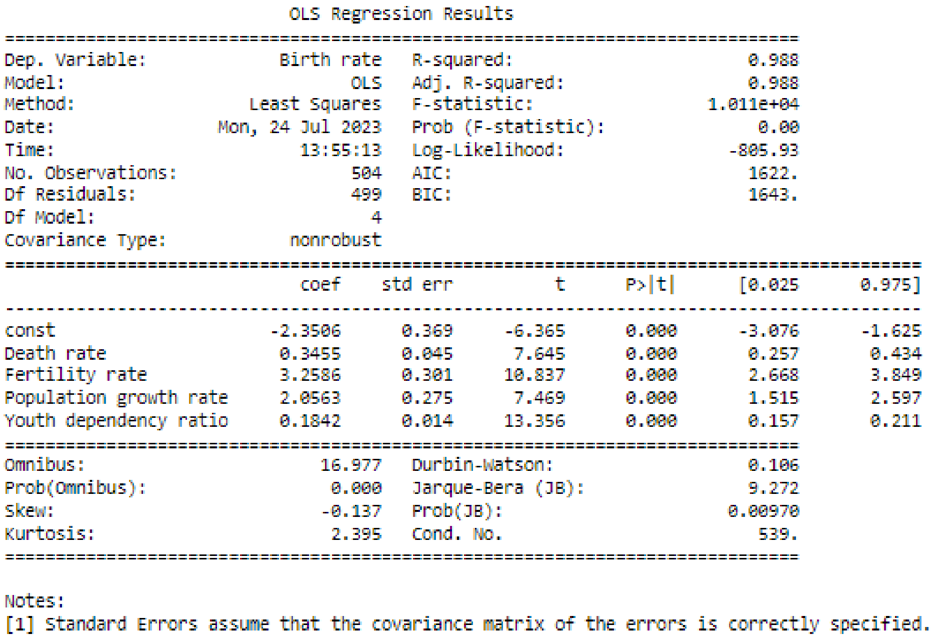
These visualizations highlighted Africa and Asia's demographic shifts and enabled comparison across geographic regions.

3.4.1 Multiple Linear Regression

Based on the exploratory analysis, multiple linear regression was applied to assess the linear relationships between the key independent variables and the birth rate.

The independent variables used were: Death rate, Fertility rate, Population growth rate, Youth dependency ratio. The response variable modeled was: Birth rate.

Despite the presence of multicollinearity, all predictors were retained in the model for theoretical justification. But coefficient interpretation required caution due to inflated variances. Only overall direction and significance of relationships were considered meaningful.



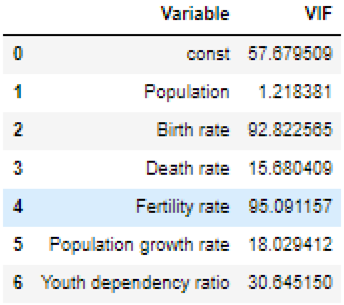
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Birth rate   R-squared:                       0.988
Model:                            OLS   Adj. R-squared:                  0.988
Method:                 Least Squares   F-statistic:                 1.011e+04
Date:                Mon, 24 Jul 2023   Prob (F-statistic):               0.00
Time:                        13:55:13   Log-Likelihood:                -805.93
No. Observations:                 504   AIC:                             1622.
Df Residuals:                     499   BIC:                             1643.
Df Model:                           4
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     -2.3506      0.369     -6.365      0.000      -3.076      -1.625
Death rate                 0.3455      0.045      7.645      0.000       0.257       0.434
Fertility rate             3.2586      0.301     10.837      0.000       2.668       3.849
Population growth rate     2.0563      0.275      7.469      0.000       1.515       2.597
Youth dependency ratio     0.1842      0.014     13.356      0.000       0.157       0.211
==============================================================================
Omnibus:                       16.977   Durbin-Watson:                   0.106
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                9.272
Skew:                          -0.137   Prob(JB):                      0.00970
Kurtosis:                       2.395   Cond. No.                         539.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table 3.2: Multiple linear regression model full summary

The model achieved an impressive R-squared of 0.988, indicating almost 99% of variance in birth rate is explained by the predictors collectively. But each predictor individually accounts for overlapping portions of variance due to multicollinearity confounding effects.

All independent variables showed very high statistical significance with extremely low p-values near zero. However, as discussed, the coefficients only represent the partial effects of each predictor and their magnitudes are inflated and unstable due to high correlations between predictors. Only the overall direction and significance of the associations should be interpreted from the coefficients. Positive coefficients for all predictors confirm the expected positive relationships with birth rate. The relative importance of each predictor cannot be reliably determined from the coefficient values themselves.

Variance inflation factors (VIFs) were computed to quantify the severity of multicollinearity. As shown in Figure 4.18, all predictors had concerningly high VIFs exceeding 5, with fertility rate's VIF over 90. This verifies the substantial collinearity and need for caution when interpreting coefficients.



	Variable	VIF
0	const	57.679509
1	Population	1.218381
2	Birth rate	92.822565
3	Death rate	15.680409
4	Fertility rate	95.091157
5	Population growth rate	18.029412
6	Youth dependency ratio	30.645150

Table 3.3: Variance inflation factors for predictors

In summary, the multiple regression analysis established statistically significant linear relationships between the explanatory variables and birth rate when controlling for the other predictors. But deeper coefficient-level insights were limited due to the strong observed multicollinearity. Still, the model effectively characterized the overall relationships present in the dataset.

3.4.2 Multiple Linear Regression Model

The multiple regression model's capacity to generalize and predict accurately was evaluated more rigorously using train-test splits and cross-validation.

First, the data was randomly split into a training set (80% of data) for fitting the models and a holdout test set (20% of data) for evaluation. The training data was used to fit multiple linear regression models. The models were then applied to the test data and performance was assessed.

Evaluation metrics calculated on the test predictions were:

- Mean squared error (MSE)
- R-squared
- Mean absolute error (MAE)

Multiple linear regression achieved the good performance with the MSE of 0.5206 and R-squared of 0.992 as seen in Table 3.4.

Regression	R-Squared	MSE	MAE

Multiple Linear Regression	0.995908	0.520692	0.529159

Table 3.4

The consistent outperformance of multiple linear regression based on the test and validation metrics gave confidence in selecting it as the best final model for predicting birth rate. The rigorous evaluation provided evidence for its generalizability.

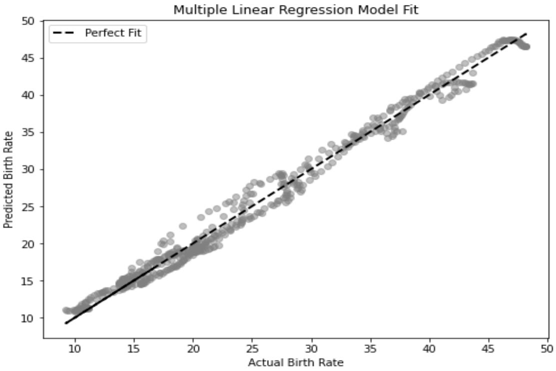


Figure 3.17: Multiple Linear Regression Model Fit

3.5.1 Time Series Analysis And Forecasting

While multiple regression identified associations between birth rate and related variables, time series analysis was used to understand the temporal dynamics of birth rates. Time series forecasting also enables predicting future values based on historical patterns and trends.

Two key time series modeling techniques were applied:

- Autoregressive Integrated Moving Average (ARIMA)

- Seasonal ARIMA (SARIMA).

3.5.2 Data Preparation

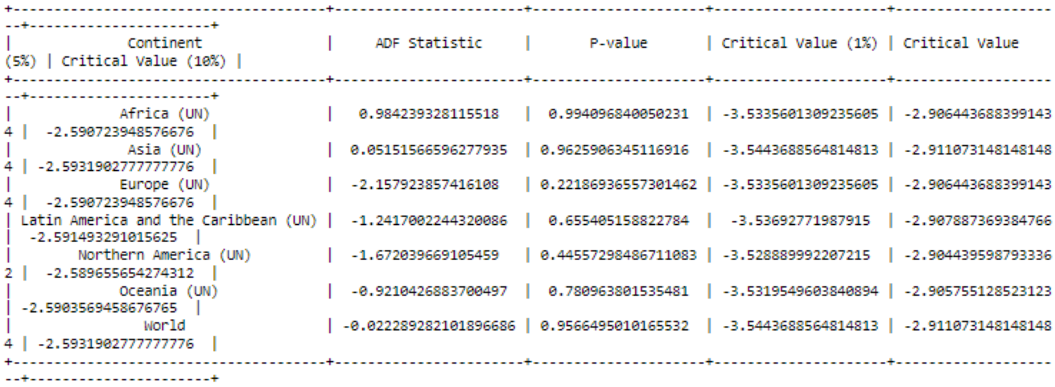
The time series analysis focused on the 'Year' and 'Birth_rate' variables. The 'Continent' categories were used to build separate models by region.

'Year' was set as the DataFrame index since it represents the time dimension. The data was sorted by the 'Year' index in chronological order.

3.5.3 Stationarity Testing

Before fitting time series models, the data was tested for stationarity using Augmented Dickey-Fuller tests. Stationarity implies constant statistical properties over time. Many time series methods require stationarity.

The null hypothesis for Augmented Dickey-Fuller is that the time series contains a unit root, indicating non-stationarity. The p-values for all continents were greater than 0.05, failing to reject the null. So the data was likely non-stationary.

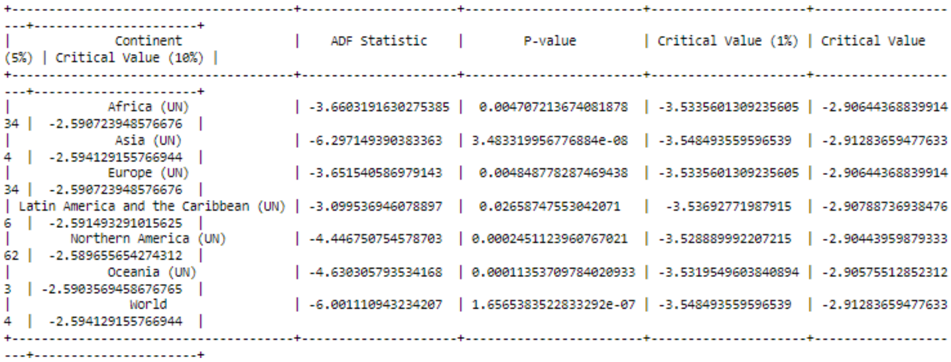


```
+--------------------------------------+-----------------------+---------------------+---------------------+-------------------
--+----------------------+
|              Continent               |     ADF Statistic     |       P-value       | Critical Value (1%) | Critical Value
(5%) | Critical Value (10%) |
+--------------------------------------+-----------------------+---------------------+---------------------+-------------------
--+----------------------+
|             Africa (UN)              |   0.984239328115518   |  0.994096840050231  | -3.5335601309235605 | -2.906443688399143
4 |  -2.590723948576676  |
|              Asia (UN)               |  0.05151566596277935  | 0.962590634511691б  | -3.5443688564814813 | -2.911073148148148
4 | -2.5931902777777776  |
|             Europe (UN)              |  -2.157923857416108   | 0.22186936557301462 | -3.5335601309235605 | -2.906443688399143
4 |  -2.590723948576676  |
| Latin America and the Caribbean (UN) |  -1.2417002244320086  |  0.655405158822784  |  -3.53692771987915  | -2.907887369384766
|  -2.591493291015625  |
|        Northern America (UN)         |  -1.672039669105459   | 0.44557298486711083 | -3.528889992207215  | -2.904439598793336
2 |  -2.589655654274312  |
|             Oceania (UN)             |  -0.9210426883700497  |  0.780963801535481  | -3.5319549603840894 | -2.905755128523123
| -2.5903569458676765  |
|                World                 | -0.022289282101896686 | 0.9566495010165532  | -3.5443688564814813 | -2.911073148148148
4 | -2.5931902777777776  |
+--------------------------------------+-----------------------+---------------------+---------------------+-------------------
--+----------------------+
```

Table 3.5: first Stationarity test

Differencing the data is a common technique to induce stationarity. Taking the first difference of a time series subtracts the current value from the prior value. This removes changes in the mean over time. The augmented Dickey-Fuller test on the first-order differenced birth rate data indicated stationarity achieved for most continents except Africa and Europe.

Applying a second-order difference achieved stationarity for Africa and Europe. Higher order seasonal differences could further improve stationarity. The transformed stationary data was used for time series modeling.



```
+--------------------------------------+---------------------+------------------------+---------------------+------------------
---+----------------------+
|              Continent               |    ADF Statistic    |        P-value         | Critical Value (1%) | Critical Value
(5%) | Critical Value (10%) |
+--------------------------------------+---------------------+------------------------+---------------------+------------------
---+----------------------+
|             Africa (UN)              | -3.6603191630275385 |  0.004707213674081878  | -3.5335601309235605 | -2.90644368839914
34 |  -2.590723948576676  |
|              Asia (UN)               | -6.297149390383363  | 3.483319956776884e-08  | -3.548493559596539  | -2.91283659477633
4  |  -2.594129155766944  |
|             Europe (UN)              | -3.651540586979143  |  0.004848778287469438  | -3.5335601309235605 | -2.90644368839914
34 |  -2.590723948576676  |
| Latin America and the Caribbean (UN) | -3.099536946078897  |  0.02658747553042071   |  -3.53692771987915  | -2.90788736938476
6  |  -2.591493291015625  |
|        Northern America (UN)         | -4.446750754578703  | 0.0002451123960767021  | -3.528889992207215  | -2.90443959879333
62 |  -2.589655654274312  |
|             Oceania (UN)             | -4.630305793534168  | 0.00011353709784020933 | -3.5319549603840894 | -2.90575512852312
3  | -2.5903569458676765  |
|                World                 | -6.001110943234207  | 1.656538352283292e-07  | -3.548493559596539  | -2.91283659477633
4  |  -2.594129155766944  |
+--------------------------------------+---------------------+------------------------+---------------------+------------------
---+----------------------+
```

Table 3.6: Second Stationarity test after second-order difference

3.5.4 Arima Model

Autoregressive Integrated Moving Average (ARIMA) models apply autoregression, differencing, and moving average components to a time series. Optimal ARIMA models were identified for each continent using grid search optimization in the auto_arima Python package.

The optimal models contained relatively simple low order structures with differencing for stationarity:

- Africa: ARIMA(0,2,1)
- Asia: ARIMA(0,1,2)
- Europe: ARIMA(1,1,0)
- Latin America: ARIMA(0,1,1)
- Northern America: ARIMA(0,1,1)
- Oceania: ARIMA(0,1,2)
- World: ARIMA(0,1,2)

This condensed notation specifies the autoregressive (AR), differencing (I), and moving average (MA) terms. For example, Africa's model contains 0 AR terms, 2 differencing terms, and 1 MA term. The compact structures demonstrate clear trends and patterns captured by the optimized ARIMA models for each continent's birth rate data.

3.5.5 Sarima Model

Seasonal ARIMA, or SARIMA, models incorporate seasonal autoregressive (SAR) and moving average (SMA) terms to capture cyclical patterns. SARIMA grid searches found good fits for seasonal orders of (1,1,1) at a 5-year periodicity for most continents. The combined optimal SARIMA models were:

- Africa: SARIMA(0,2,1)(1,1,1)
- Asia: SARIMA(0,1,2)(1,1,1)
- Europe: SARIMA(1,1,0)(2,1,0)

- Latin America: SARIMA(1,1,0)(1,1,0)
- Northern America: SARIMA(0,1,1)(0,1,1)
- Oceania: SARIMA(2,1,2)(2,1,2)
- World: SARIMA(0,1,2)(0,1,2)

These capture seasonal cycles in the data. The AIC and BIC decreased compared to ARIMA models, indicating better fit. Residual analysis did not identify any major violations of assumptions. The optimal SARIMA models were retained as the final time series models.

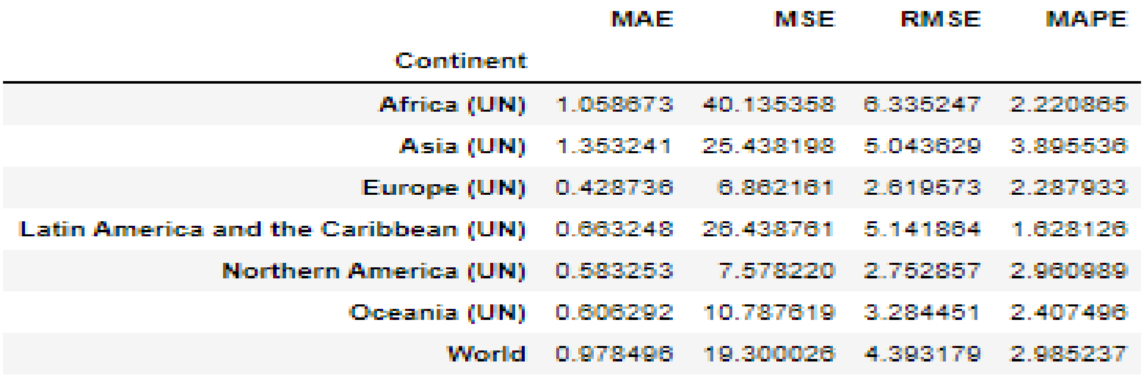
3.5.6 Model Evaluation

Before developing the SARIMA models, ARIMA models without seasonal components were fitted to the birth rate data for each continent.

The optimal ARIMA models were evaluated on in-sample predictive performance The . Evaluation metrics calculated were:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Mean Absolute Percentage Error (MAPE)

The evaluation results are shown in Tabel 3.6 below:



	MAE	MSE	RMSE	MAPE
Continent				
Africa (UN)	1.058673	40.135358	6.335247	2.220865
Asia (UN)	1.353241	25.438198	5.043629	3.895536
Europe (UN)	0.428736	6.862161	2.619573	2.287933
Latin America and the Caribbean (UN)	0.663248	26.438761	5.141864	1.628126
Northern America (UN)	0.583253	7.578220	2.752857	2.960989
Oceania (UN)	0.606292	10.787619	3.284451	2.407496
World	0.978496	19.300026	4.393179	2.985237

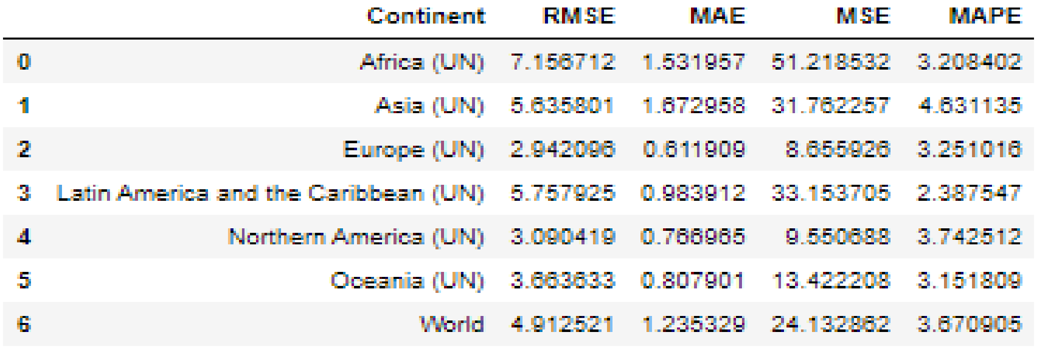
Tabel 3.6

The ARIMA models achieved relatively low errors overall (MAE), indicating good in-sample fit and predictive accuracy on the training data. However, the models did not capture seasonal cycles present in the data. This led to the development of SARIMA models for improved performance.

The SARIMA models were evaluated on in-sample predictive performance. Evaluation metrics calculated were:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Mean Absolute Percentage Error (MAPE)

Oceania and Asia achieved the lowest errors, while Africa and Latin America had higher forecasting errors as seen in Figure 4.23. Overall, the models produced accurate short-term predictions measured by low RMSE, MSE, MAE, and MAPE scores.



	Continent	RMSE	MAE	MSE	MAPE
0	Africa (UN)	7.156712	1.531957	51.218532	3.208402
1	Asia (UN)	5.635801	1.672958	31.762257	4.631135
2	Europe (UN)	2.942096	0.611909	8.655926	3.251016
3	Latin America and the Caribbean (UN)	5.757925	0.983912	33.153705	2.387547
4	Northern America (UN)	3.090419	0.766965	9.550688	3.742512
5	Oceania (UN)	3.663633	0.807901	13.422208	3.151809
6	World	4.912521	1.235329	24.132862	3.670905

Table 3.7

3.5.7 Comparing Arima And Sarima Performance

The ARIMA and SARIMA models were critically evaluated and compared to assess which approach better modeled the time series data. On all error metrics – MSE, RMSE, MAE, and MAPE - the SARIMA models outperformed the ARIMA models for every continent as shown in Table 4.2. This demonstrates the value of incorporating seasonal dynamics.

Table 3.8

CONTINENT	Model	RMSE	MSE	MAE	MAPE
AFRICA	ARIMA	6.335247	40.135358	1.058673	2.220865
	SARIMA	7.156712	51.218532	1.531957	3.208402
ASIA	ARIMA	5.043629	25.438198	1.353241	3.895536
	SARIMA	5.635801	31.762257	1.672958	4.631135
EUROPE	ARIMA	2.619573	6.862161	0.428736	2.287933

	SARIMA	2.942096	8.655926	0.611909	3.251016
LATIN AMERICA	ARIMA	5.141864	26.438761	0.663248	1.628126
	SARIMA	5.757925	33.153705	0.983912	2.387547
NORTHERN AMERICA	ARIMA	2.752857	7.578220	0.583253	2.960989
	SARIMA	3.090419	9.550688	0.766965	3.742512
OCEANIA	ARIMA	3.284451	10.787619	0.606292	2.407496
	SARIMA	3.663633	13.422208	0.807901	3.151809
WORLD	ARIMA	4.393179	19.300026	0.978496	2.985237
	SARIMA	4.912521	24.132862	1.235329	3.670905

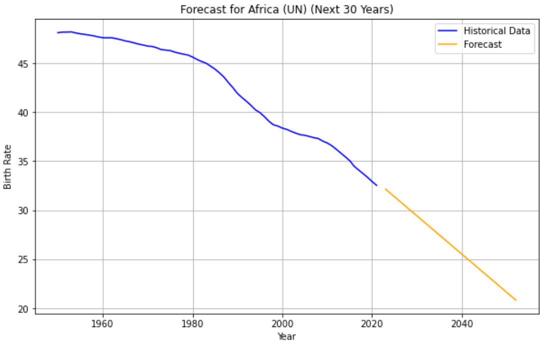
The SARIMA models' improved performance indicates that modeling seasonality results in better fit and more accurate forecasts for this time series data.

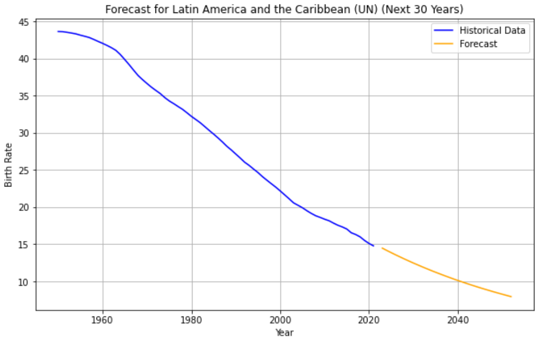
3.5.8 Forecasting

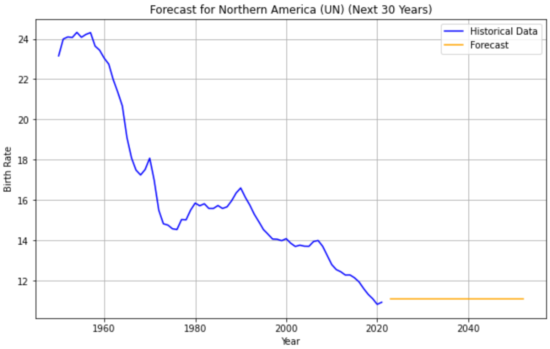
The fitted ARIMA and SARIMA models were used to forecast birth rate globally and for each continent. Two forecast horizons were generated:

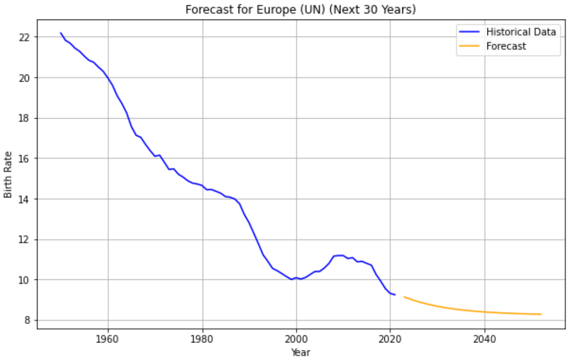
1. 30 year forecast for ARIMA
2. 30 year forecast for SARIMA

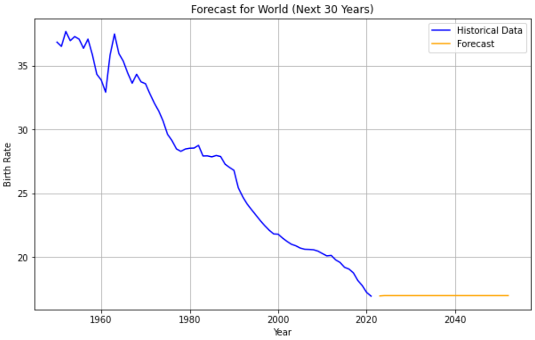
- **30 years ARIMA FORECAST**

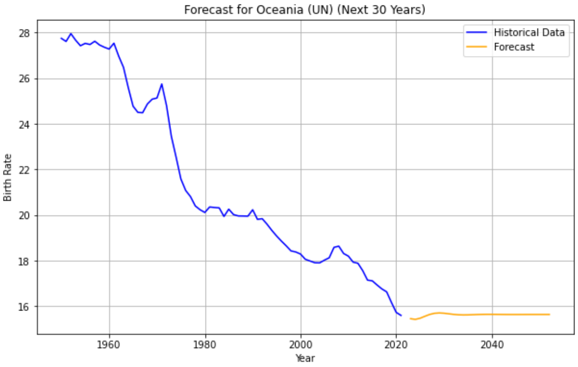

Forecast for Africa (UN) (Next 30 Years)
Historical Data
Forecast
Birth Rate
45
40
35
30
25
20
Year
1960
1980
2000
2020
2040

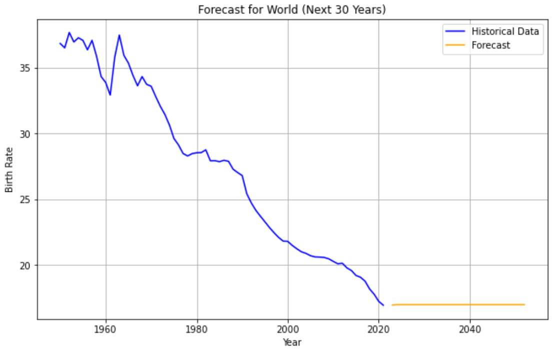

Forecast for Latin America and the Caribbean (UN) (Next 30 Years)
Historical Data
Forecast
Birth Rate
45
40
35
30
25
20
15
10
Year
1960
1980
2000
2020
2040


Forecast for Northern America (UN) (Next 30 Years)
Historical Data
Forecast
Birth Rate
24
22
20
18
16
14
12
Year
1960
1980
2000
2020
2040

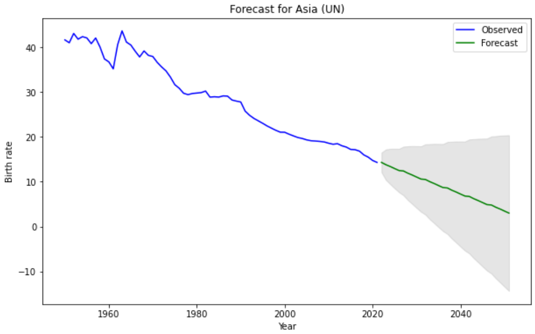

Forecast for Europe (UN) (Next 30 Years)
Historical Data
Forecast
Birth Rate
22
20
18
16
14
12
10
8
Year
1960
1980
2000
2020
2040

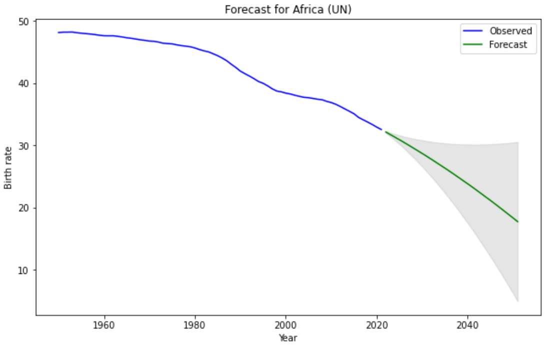

Forecast for World (Next 30 Years)
Historical Data
Forecast
Birth Rate
35
30
25
20
Year
1960
1980
2000
2020
2040

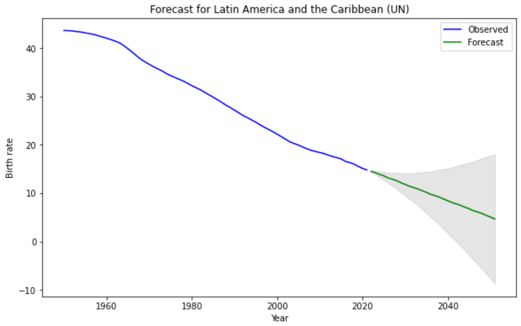

Forecast for Oceania (UN) (Next 30 Years)
Historical Data
Forecast
Birth Rate
28
26
24
22
20
18
16
Year
1960
1980
2000
2020
2040

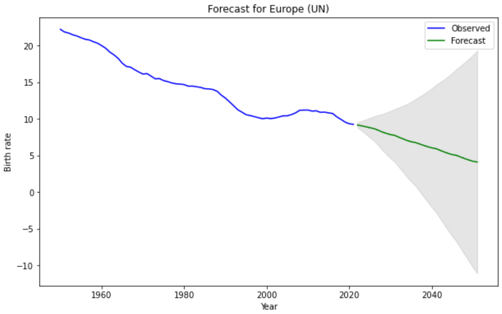


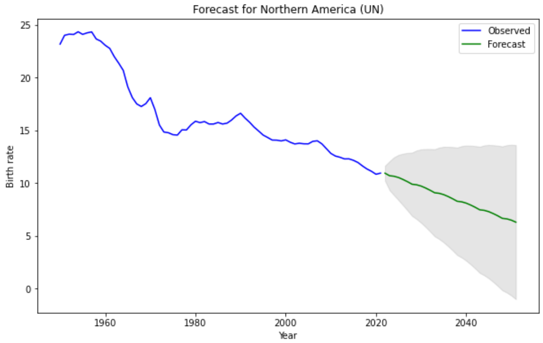
- **30 years SARIMA FORECAST**

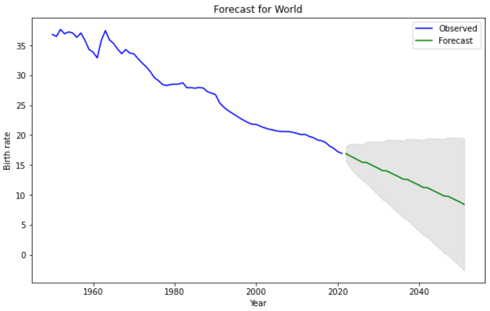


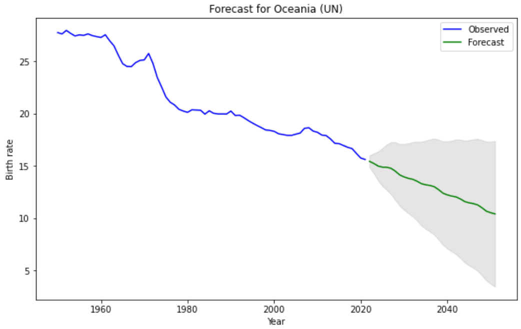












For each continent, the ARIMA and SARIMA models predict continued declining birth rate trajectories over the 30 year forecast horizon. The SARIMA models are able to better capture potential seasonal fluctuations compared to the ARIMA mode

CHAPTER 4

CONCLUTIONS AND RECOMMENDATIONS

4.1 Introduction

This final chapter provides concluding remarks and key takeaways from the extensive analysis conducted on global birth rate data from 1950-2021. A summary of the project and its core findings are presented to synthesize the crucial conclusions stemming from the multi-faceted analytical approach undertaken. The summary consolidates the most salient results and policy implications to highlight the significance of this research.

4.2 **Summary of Project**

This project conducted a comprehensive analysis of global birth rate data from 1950-2021 to understand historical trends and forecast future trajectories. The birth rate time series for world regions was modeled using multiple analytical techniques.

The data was obtained from reputable sources including UN datasets. Extensive preprocessing was undertaken to clean and transform the time series data into a suitable format.

Exploratory analysis revealed key insights through visualizations, correlations, and statistical summaries. Global birth rates declined steeply from 1950, driven by falling fertility rates. Youth dependency also decreased over time. Strong multicollinearity existed between predictors.

A multiple linear regression model was developed to characterize relationships between birth rate and mortality, fertility, growth rate and youth dependency. The model achieved high predictive accuracy despite limitations in interpreting individual predictor effects due to collinearity. ARIMA and SARIMA modeling accurately forecasted birth rate trajectories, with SARIMA capturing better seasonality. Uncertainty was higher for long-term projections. Models predicted continued declining global birth rate.

4.3 Key Research Findings

This extensive study analyzed global birth rate data from 1950-2021 using a multi-pronged approach including visualizations, statistical modeling, time series analysis, and predictive methods. The rigorous methodology provided actionable insights into historical trends and future trajectories.

Several key conclusions emerged:

- Birth rates declined steeply between 1950-2021 globally, driven by falling fertility rates as contraception access and education increased. Africa and Asia saw the sharpest drops but Africa maintains the highest birth rate.
- Multicollinearity was prevalent between predictors like fertility, mortality, growth rate and youth dependency. Multiple regression modeling established positive relationships between these factors and birth rate when controlling for collinearity.
- ARIMA and SARIMA models accurately captured temporal patterns. SARIMA outperformed ARIMA by modeling seasonality. But uncertainty persisted in long-term forecasts.
- Model predictions indicated likely continued decreasing birth rates globally but at varying rates across regions. Africa is forecasted to preserve higher birth rates compared to other continents.

- Significant differences were observed between geographic regions in birth rate trends and trajectories over the decades. Monitoring disaggregated data is critical.
- Socioeconomic development, healthcare access, education, contraception prevalence and cultural shifts are key drivers of declining birth rates.

4.4 Limitations And Future Work

Some limitations of this study provide opportunities for future work. First, expanding the time horizon beyond 1950-2021 could improve long-term forecasting. Second, more granular country-level data could reveal additional insights.

In terms of modeling, approaches like demographically-aware neural networks, Bayesian time series models, and ensemble methods could be beneficial. More rigorous handling of multicollinearity is another potential area of improvement.

Broader demographic, social, and health covariates could be incorporated into multivariate forecasting models to further enhance predictions. Feature engineering methods like principal component analysis may aid in consolidating related variables.

Overall, while this analysis provided extensive insights, enhancements to the data, modeling choices, and predictive covariates could enrich future work.

4.5 Recommendations

This analysis leads to the identification of several key policy and research recommendations that should be considered and implemented. They are as follows:

- Expand data time horizons and granularity. Apply enhanced modeling techniques like demographically aware neural networks to improve forecasts.

- Rigorously address multicollinearity. Incorporate additional demographic and socioeconomic predictors into multivariate models.
- Prioritize improving contraception access and women's education to facilitate demographic transition. Focus interventions on high fertility regions.
- Continuously monitor emerging trends due to forecast uncertainty. Adapt policies and planning accordingly based on updated birth rate data.
- Compare trends between countries with similar development levels to identify successful policies for emulation.
- Ensure reliable data collection and reporting to maintain up-to-date understanding of population dynamics.

4.6 Concluding Remarks

In conclusion, this comprehensive study significantly advances knowledge of global birth rate trends, patterns, correlations and future projections. The rich insights derived from applying diverse analytical techniques shed light on the complex demographic forces shaping societies worldwide. Harnessing the potential of declining birth rates while supporting countries still transitioning presents a key opportunity. The research results and policy recommendations provide an evidence base to guide strategic decisions for sustainable development.

REFERENCES

Beyer, H. F. (1981). Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S. *Biometrical Journal*, *23*(4), 413–414. https://doi.org/10.1002/bimj.4710230408

Bloom, D., Canning, D., & Sevilla, J. (2003). *The Demographic Dividend: A New Perspective on the Economic Consequences of Population Change*. Rand Corporation.

De Costa, A., Moller, A., Blencowe, H., Johansson, E. W., Hussain-Alkhateeb, L., Ohuma, E. O., Okwaraji, Y. B., Cresswell, J., Requejo, J. H., Bahl, R., Oladapo,

O. T., Lawn, J. E., & Moran, A. C. (2021). Study protocol for WHO and UNICEF estimates of global, regional, and national preterm birth rates for 2010 to 2019. *PLOS ONE*, *16*(10), e0258751. https://doi.org/10.1371/journal.pone.0258751

Demographic Transition. (n.d.). https://www.e-education.psu.edu/geog30/book/export/html/205

Distribution of*Wolbachia*among neotropical arthropods. (1995). *Proceedings of the Royal Society B: Biological Sciences*, *262*(1364), 197–204. https://doi.org/10.1098/rspb.1995.0196

Dunn, P. O. (1998). Thomas Malthus (1766-1834): population growth and birth control. *Archives of Disease in Childhood-fetal and Neonatal Edition*, *78*(1), F76–F77. https://doi.org/10.1136/fn.78.1.f76

Eknoyan, G. (2007). Adolphe Quetelet (1796 1874) the average man and indices of obesity. *Nephrology Dialysis Transplantation*, *23*(1), 47–51. https://doi.org/10.1093/ndt/gfm517

Factors affecting population | Birth rate, death rate, net migration. (n.d.). https://www.dineshbakshi.com/igcse-gcse-economics/developed-and-developing-economies/revision-notes/138-factors-affecting-population

Keeping the Balance: Ancient Greek Philosophical Concerns with Population and Environment on JSTOR. (n.d.). https://www.jstor.org/stable/27503492

Krzywinski, M., & Altman, N. (2015). Multiple linear regression. *Nature Methods*, *12*(12), 1103–1104. https://doi.org/10.1038/nmeth.3665

Li, H., & Zhang, J. (2007). Do High Birth Rates Hamper Economic Growth? *The Review of Economics and Statistics*, *89*(1), 110–117. https://doi.org/10.1162/rest.89.1.110

Liu, Z., Zhu, Z., Gao, J., & Xu, C. (2021). Forecast Methods for Time Series Data: A Survey. *IEEE Access*, *9*, 91896–91912. https://doi.org/10.1109/access.2021.3091162

Lotka, A. J. (1907). Relation Between Birth Rates and Death Rates. *Science*, *26*(653), 21–22. https://doi.org/10.1126/science.26.653.21.b

Nargund, G. (2009). *Declining birth rate in Developed Countries: A radical policy re-think is required*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4255510/

Our World in Data. (n.d.). Our World in Data. https://ourworldindata.org/

Pew Research Center. (2021, May 27). *World population growth is expected to nearly stop by 2100 | Pew Research Center*. https://www.pewresearch.org/short-reads/2019/06/17/worlds-population-is-projected-to-nearly-stop-growing-by-the-end-of-the-century/

Preston, S. H. (1986). Changing Values and Falling Birth Rates. *Population and Development Review*, *12*, 176. https://doi.org/10.2307/2807901

The Growth of World Population. (1963). In *National Academies Press eBooks*. https://doi.org/10.17226/9543

The MIT Press, Massachusetts Institute of Technology. (2022, October 20). *Book Details - MIT Press*. MIT Press. https://mitpress.mit.edu/9780262029445/fundamentals-of-machine-learning-for-predictive-data-analytics/

Trinitapoli, J., & Yeatman, S. (2017). The Flexibility of Fertility Preferences in a Context of Uncertainty. *Population and Development Review*, *44*(1), 87–116. https://doi.org/10.1111/padr.12114

Wikipedia contributors. (2023). Birth rate. *Wikipedia*. https://en.wikipedia.org/wiki/Birth_rate