JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

GLOBAL LIFE EXPECTANCY PREDICTION USING NONSTATIONARY TIME SERIES MODELS

[1]Ambati Sushmitha, [2]Ravipalli Shreya, [3]Donakanti Sharmila, [4]Dr. M.V. Krishna Rao
[1,2,3]B. Tech Student of Department of CSE (Data Science), Institute of Aeronautical Engineering.
[4]Professor, Department of CSE (Data Science), Institute of Aeronautical Engineering, Hyderabad, India

Abstract: Life expectancy, a critical measure of public health and social progress, indicates the average lifespan of a population and has become essential for shaping healthcare policies, economic planning, and demographic research. Predicting life expectancy trends allows stakeholders to prepare for future population dynamics, allocate resources, and implement necessary interventions effectively. Early Life Expectancy models primarily relied on mortality rates within specific populations; however, advancements in forecasting now emphasize the importance of additional factors—such as education, economic conditions, and healthcare access—in accurately projecting life expectancy trends. In this study, life expectancy values of several countries are forecasted utilizing: Principal Component Analysis (PCA), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) with different differencing order. Then it compares the performance of these methods to understand predicting future life expectancy trends.

***Index Terms* - Life expectancy prediction, Principal Component Analysis (PCA), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), time series analysis, demographic forecasting.**

I. INTRODUCTION

Forecasting life expectancy plays a crucial role in understanding public health policies, demographic research, and planning social services. By predicting life expectancy, governments and healthcare industry can improve outcomes by having better resources, long-term strategies and understand future healthcare demands. Time series models have become essential tools for predicting complex trends, such as life expectancy. This research applies Principal Component Analysis (PCA), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA) models to forecast life expectancy. These models help identify patterns in historical data and project future trends. The study aims to evaluate the effectiveness of these models in life expectancy prediction and compare their strengths and limitations.

Life expectancy is a key indicator to manage population's health and well-being. It reflects the average lifespan a person is expected to live, based on current mortality patterns. Accurate predictions of life expectancy are essential for applications such as public health planning, social security systems, and healthcare resource management. Given the various factors influencing life expectancy—such as healthcare access, socio-economic conditions, and lifestyle changes—accurate and reliable forecasting models are crucial for future planning.

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of datasets. It transforms the original data into principal components— uncorrelated variables that capture the most significant variation in the data. This simplifies the model, improving interpretability and efficiency without sacrificing accuracy. In this study, PCA is employed to preprocess life expectancy data, helping reduce noise and enhance the performance of the time series forecasting models by focusing on the most relevant factors.

Time series models analyze historical data to predict future outcomes, making them suitable for life expectancy forecasting, which requires understanding both short-term fluctuations and long-term trends. While effective for straightforward patterns. Autoregressive Integrated Moving Average (ARIMA) models extend ARMA by introducing a differencing step to address nonstationary data, where trends and patterns change over time. ARIMA is particularly effective for long-term forecasts, as it stabilizes evolving data and improves accuracy. Autoregressive Moving Average (ARMA) models incorporate both autoregressive and moving average components, combining historical data relationships with residual error corrections to enhance forecasting accuracy. ARMA is often used for mid-range forecasting due to its balanced approach to capturing trends and noise.

The comparison of ARMA, ARIMA, and PCA models provides valuable insights into their performance under different conditions: ARMA improves forecasting accuracy by incorporating both past data and error corrections, making it ideal for mid-range predictions. ARIMA is well-suited for long-term forecasting, especially when dealing with non-stationary data that exhibits evolving trends over time. PCA enhances model efficiency and interpretability by simplifying the dataset, which can boost the performance of time series models, particularly in complex datasets. By comparing these models, this study underlies a framework for understanding of the strengths and limitations of each approach, offering valuable insights for researchers and policymakers involved in demographic forecasting and public health planning.

II. Literature Review

In his paper, Schöley, Jonas, et al. [1] analyze the life expectancy losses caused by the COVID-19 pandemic across 29 countries, including the U.S., Chile, and much of Europe, beginning in 2020. By assessing changes in mortality rates across various age groups, we explore how the pandemic affected different populations. Our findings show a divergence in the impact of COVID-19 by 2021. While Western European countries experienced a recovery in life expectancy, other regions continued to suffer significant losses. This variation highlights the uneven effects of the pandemic on global mortality trends.

Woolf, Steven H., et al. [2] examines the significant decrease in U.S. life expectancy in 2020 due to COVID-19, with larger impacts on Hispanic and non-Hispanic populations compared to non-Hispanic Whites. The U.S. experienced a 1.87-year drop in life expectancy, while 21 peer countries saw an average decrease of only 0.58 years. The findings emphasize the pandemic's disproportionate effect on racial minorities in the U.S., highlighting the need to address systemic health inequities.

Perumalsamy, et al. [5] examines the use of AI in improving mortality and longevity predictions for the life insurance industry, surpassing traditional actuarial methods. AI offer greater accuracy by analyzing complex, diverse data, leading to more personalized and dynamic risk assessments. Challenges include data quality, model transparency, and regulatory adaptation. The paper concludes that AI has the potential to significantly enhance underwriting and risk management in life insurance.

Bali, Vikram, et al. [6] explores the limitations of traditional life expectancy (LE) models, which primarily focused on mortality data. It highlights the need to incorporate additional factors like education, health, and economic conditions to improve predictions. The authors use machine learning algorithms to enhance the accuracy.

Ye, Wenjing, et al. [7] analyze data from 213 COVID-19 patients in Wuhan to identify different patient subgroups using principal component and cluster analysis. It highlights three distinct clusters based on age, immune function, and disease severity, with Cluster 1 showing the highest mortality and severe outcomes. The study emphasizes that factors like immune function and albumin levels, rather than age alone, are crucial for predicting disease severity and guiding treatment strategies.

Choubey, Dilip K., et al. [8] focus on developing an efficient diagnostic system for early detection of diabetes using machine learning techniques. It compares two approaches: one with direct classification methods (Adaboost, CVR, KNN, RBFN) and another with feature reduction methods (PCA, LDA) followed by the same classifiers, applied to both the Pima Indian Diabetes Dataset and a local dataset. The study finds that PCA combined with CVR offers the best performance, and concludes that feature reduction improves accuracy, reduces computation time, and can be applied to other medical diagnoses.

He, Yunting, et al. [11] introduces a new "epidemic evaluation index" (EEI) using a seven-day moving average of daily new cases, combined with neural networks like CNNs, to predict COVID-19's peak. Validated with SARS data, the EEI showed China's peak in early February 2020, while many countries hadn't peaked by mid-April. This method helps inform decisions on resuming normal activities.

Singh, Ram Kumar, et al. [12] analyzes COVID-19 spread in the top 15 countries as of April 2020, comparing cases, deaths, and recoveries. Using an ARIMA model, it predicts the pandemic's trajectory over two months and identifies trends in spread, recovery, and death rates across regions.

Perone, Gaetano, et al. [14] used the ARIMA model to forecast the trend of COVID-19 in Italy from February 20 to April 4, 2020, and outlined how future trends and inflection points of the epidemic will be identified. This method can explain how much social and economic impact it would bring to the country.

Awariefe, C., et al. [19] forecast the survival rates of Nigeria's population aged 65 and older, assessing trends up to 2030. Using time series data, various ARIMA models were applied to identify the best fit. The results indicate that survival rates for both males and females are expected to increase in the coming years, assuming no significant epidemics that could raise mortality rates.

Ayele, et al. [20] examines the COVID-19 pandemic's impact on global health and the economy, noting the significant rise in confirmed cases. It utilizes an Autoregressive Moving Average (ARMA) model to forecast daily incidence rates based on early data. The analysis identifies the best-fit models for daily deaths and confirmed cases, highlighting the urgent need for effective public health interventions and adherence to WHO guidelines to manage the outbreak.

III. METHODOLOGY

With clean data, three forecast models were then applied to this reduced dataset:

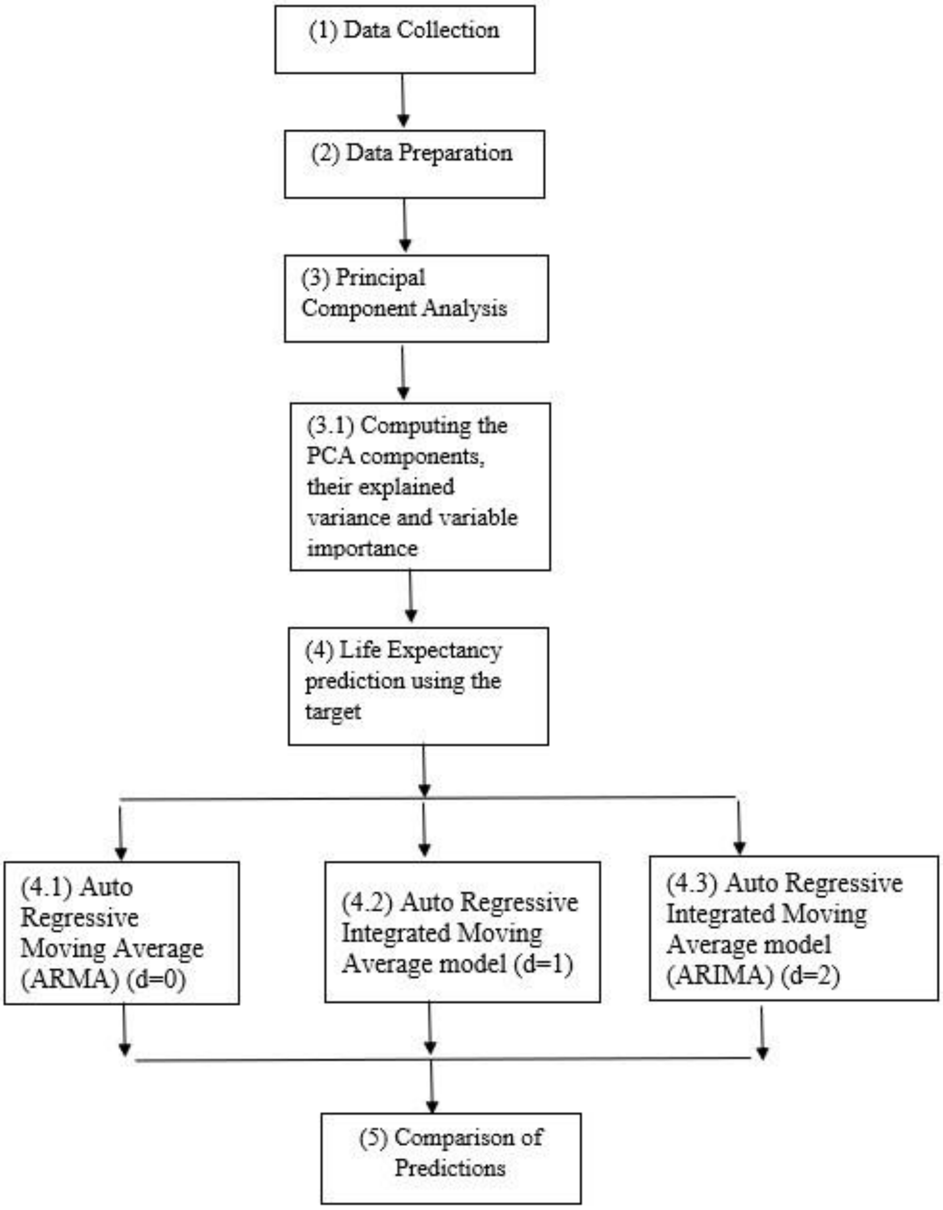

(1) Data Collection
(2) Data Preparation
(3) Principal Component Analysis
(3.1) Computing the PCA components, their explained variance and variable importance
(4) Life Expectancy prediction using the target
(4.1) Auto Regressive Moving Average (ARMA) (d=0)
(4.2) Auto Regressive Integrated Moving Average model (d=1)
(4.3) Auto Regressive Integrated Moving Average model (ARIMA) (d=2)
(5) Comparison of Predictions

Figure 1: step by step process of life expectancy analysis.

3.1 Data Preparation

The data used in this study was sourced directly from the website of the United Nations and from the repository of the World Health Organization. This study covered a time span of years from 1960 to 2022. The dataset had several key columns including country name, year, life expectancy, GDP per capita (current US$), Immunization, DPT (% of children ages 12-23 months), Immunization, HepB3 (% of one-year-old children), Immunization, measles (% of children ages 12-23 months), Adult Mortality rate female (per 1,000 female adults), Adult Mortality rate male (per 1,000 male adults), Population, Total alcohol consumption per capita, Number of under-five deaths, Infant Mortality Rate (for male, female and Both Sexes) and a categorical variable labeled "Status" denoting whether a country was declared as developed or developing. This extensive dataset was the basis for the analysis and prediction tasks.

3.2 Data Preprocessing

Preprocessing of the data for modeling involved extensive steps. Missing values in the dataset were imputed using imputation techniques. Numerical columns were handled using mean or median imputation, while categorical variables, such as the "Status" column, were treated using mode imputation. To enable machine learning algorithms to process the data effectively, categorical variables were converted into numerical representations using one-hot encoding. Furthermore, numerical features were normalized to ensure all variables contributed equally during the modeling process, avoiding biases caused by differing scales.

3.3 Data Cleaning

The next stage involved data cleaning and dimensionality reduction, achieved through the application of Principal Component Analysis (PCA). It helped to reduce the dimensionality of the dataset but retain the information that was of most importance. PCA identified principal components that could explain 95% of cumulative variance in data, ensuring only the most important features were carried forward for analysis. This did not only save computation time but also improved interpretability of later predictive models.

3.4 Data Analysis

In the area of modeling for life expectancy prediction, time series models that are statistical in nature, such as ARIMA, ARMA and are used to model trends, cycles, and randomness present in historical data. These statistical models can forecast based on prior values of life expectancy and have flexibility in complexity and sophistication in regard to the temporal data.

3.4.1 ARMA (Auto Regressive Moving Average)

The ARMA (Autoregressive Moving Average) model is built on two components; the Autoregressive (AR) portion which forecasts future values from past observations, and the Moving Average (MA) section which captures the relation between current value and past forecast errors. ARMA is a model of stationary time series, which means that the statistical properties of the data remain unchanged through time. The ARMA is defined as

$$ARMA(p,q): Y_t = c + \sum_{i=1}^{p} \emptyset_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \in_{t-i} + \in_t \qquad (3.4.1)$$

3.4.2 ARIMA (Auto Regressive Integrated Moving Average)

The three elements are an autoregressive (AR), integrated (I) that means taking the difference to achieve stationarity, and a moving average (MA)[14]. In the ARIMA model, the ARMA is extended by inclusion of the differencing step for nonstationary data. The ARIMA can be defined using three parameters (p, d, q) as p: count of AR terms d: degree of differencing to make the series stationary q: number of MA terms. General ARIMA model
This equation is as follows;

$$\Delta^D_{y_t} = \sum_{i=1}^{p} \emptyset_i \Delta^D_{y_t} + \sum_{j=1}^{q} \emptyset_j \in_{t-j} + \sum_{m=1}^{M} \beta_m X_{mt} + \in_t \qquad (3.4.2)$$

3.5 Evaluation

Next, the performance metrics of the models were computed and compared.

IV. RESULTS AND DISCUSSION

The life expectancy dataset encompasses information of 2,938 records and 22 variables. It uses Principal Component Analysis (PCA) in order to reduce the dimensionality of the dataset and, as additional information, the most significant factors causing variance in life expectancy between countries. Some preprocessing steps involved in this process were missing value imputation by making use of the median strategy and standardizing the dataset to make all the features comparable on a basis. Before performing PCA, the correlation heatmap depicted in Figure 2 was obtained to evaluate interdependencies among the original variables. From the heatmap, employing different colors indicating varying levels of correlation, many strong interdependencies were identified between the features.

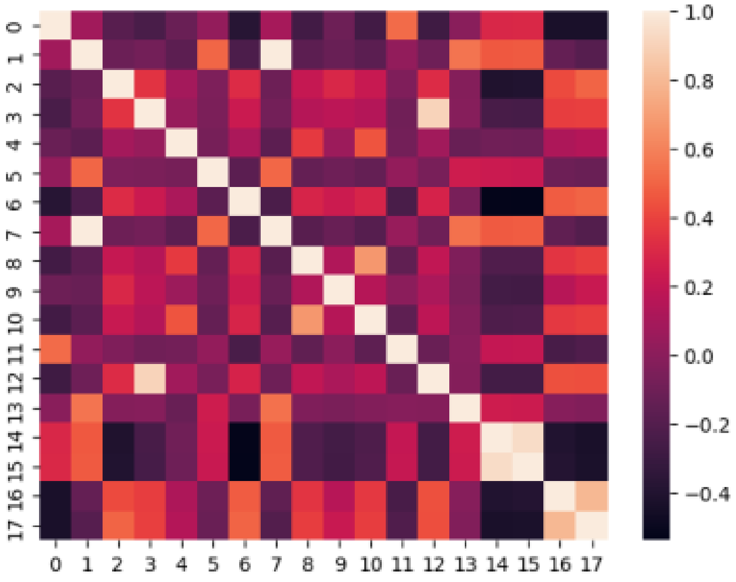


Figure 2 Correlation of Data

After applying standardization to the data set, PCA was conducted to obtain principal components. The cumulative explained variance plot Figure 3 is elaborated below. The number of components that would retain 95% of the variance was determined with the help of running of the cumulative explained variance. It showed that 13 components would be a threshold level to achieve this.

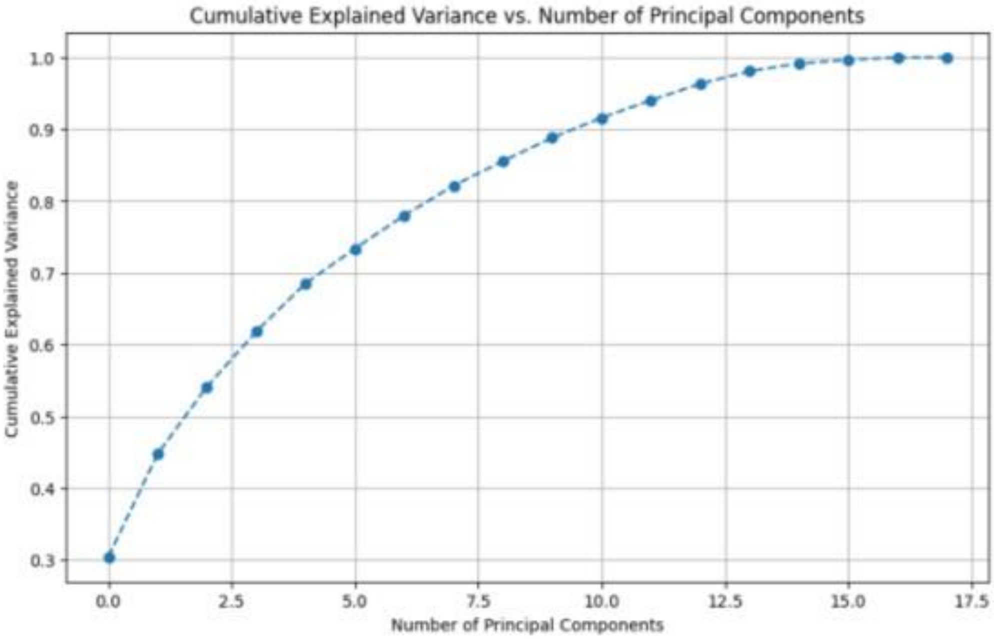


Figure 3 Cummulative Explained vs Number of Principal Components

However, the first component, now famously called PC1, alone accounts for 30.44%, followed by the second component accounting for 14.33%. Together, the first two components explain nearly 45% of the variance and hence very important to grasp the structure of the dataset. Thus, in such a way, dimensionality reduction enables the data handling while keeping most of the information significant to the data. After performing PCA, the resulting correlation heatmap (Figure 4) shows no correlation between the principal components.

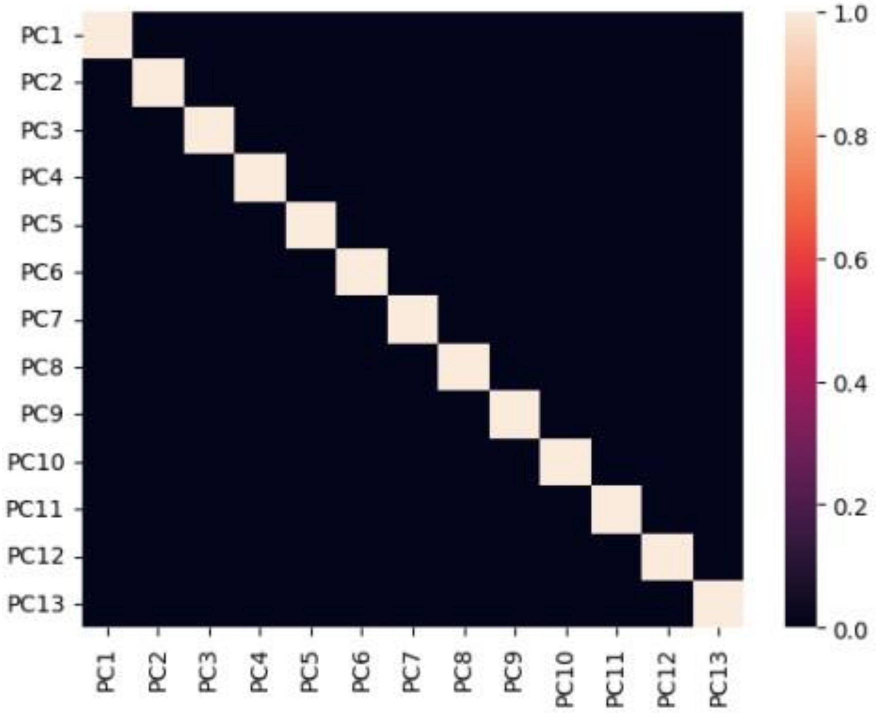


Figure 4 Correlation of PCA Data

Figure 5 aggregates the loadings of all original variables on to the principal components were calculated to understand the contributions of individual variables. For example, the loadings of , income composition of resources , thinness 5-9, schooling, and thinness 1-19 years figure very prominently on the first principal component-meaning that these three variables are playing the key role in life expectancy. Other major contributors across the first three components Polio, Diphtheria, and infant mortality. Table 1 aggregates the loadings of all original variables on to the principal components were calculated to understand the contributions of individual variables. For example, the loadings of, income composition of resources , thinness 5-9, schooling, and thinness 1-19 years figure very prominently on the first principal component-meaning that these three variables are playing the key role in life expectancy. Other major contributors across the first three components Polio, Diphtheria, and infant mortality.

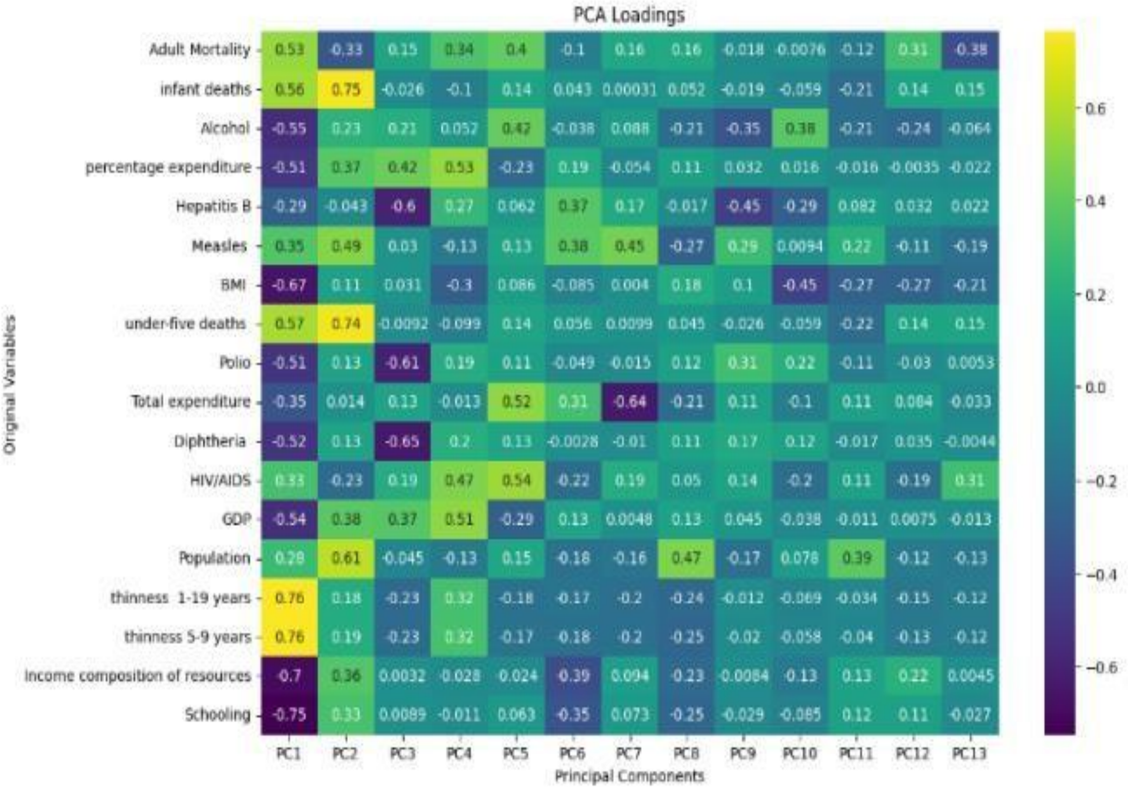
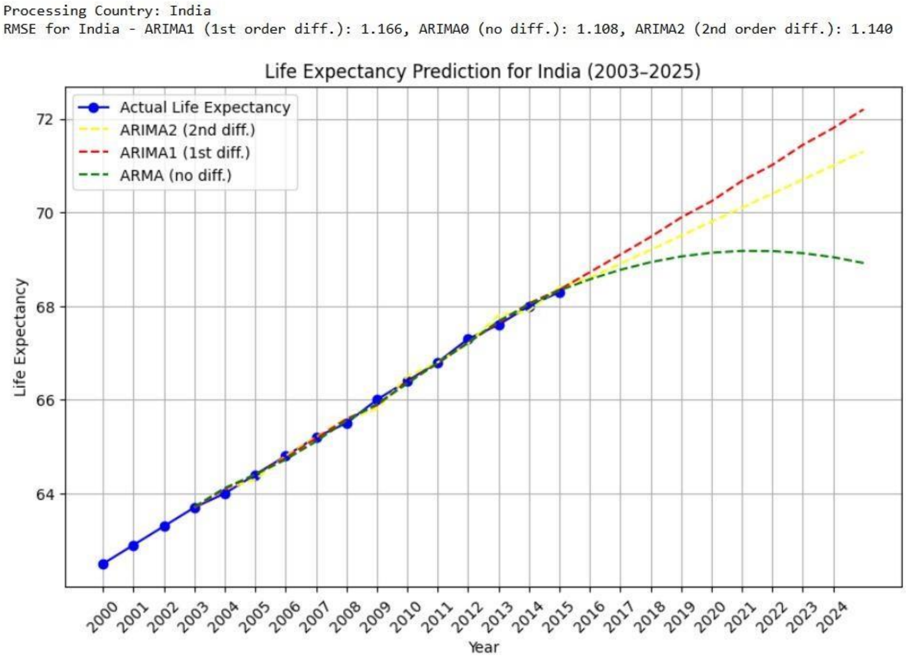


Figure 5 PCA Loadings

After performing of PCA, the analysis of life expectancy trends for various nations between 2000 and 2025 was conducted using time series forecasting including ARIMA with first-order differencing (ARIMA1), ARIMA with second-order differencing (ARIMA2), and ARMA with no differencing (ARIMA0).The actual life expectancy values from historical data are plotted alongside these predictions to evaluate the models' performance.

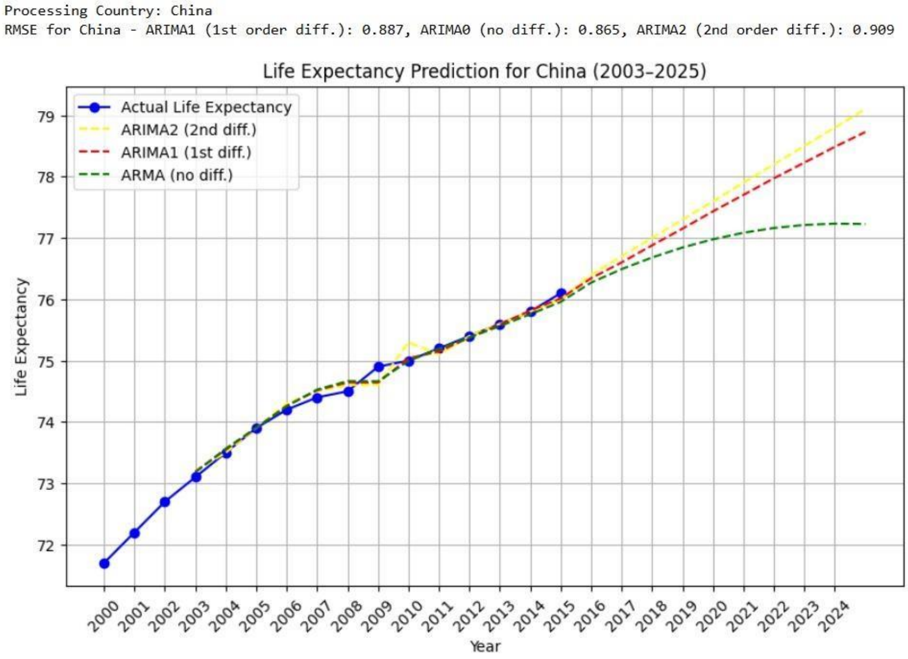
For India(figure 6), all three models captured the upward trend in life expectancy, with ARIMA1 predicting a stronger increase, while ARMA and ARIMA2 showed more conservative projections. Similarly, in China, the actual trend exhibited a steady rise, with ARIMA1 and ARIMA2 projecting continued growth, whereas ARMA suggested a more stable trajectory. These results indicate that ARMA provides a more generalized fit, whereas ARIMA models emphasize recent trends, which may be useful when dealing with non-stationary data.



```
Processing Country: India
RMSE for India - ARIMA1 (1st order diff.): 1.166, ARIMA0 (no diff.): 1.108, ARIMA2 (2nd order diff.): 1.140
```

Figure 6 Predictios of India

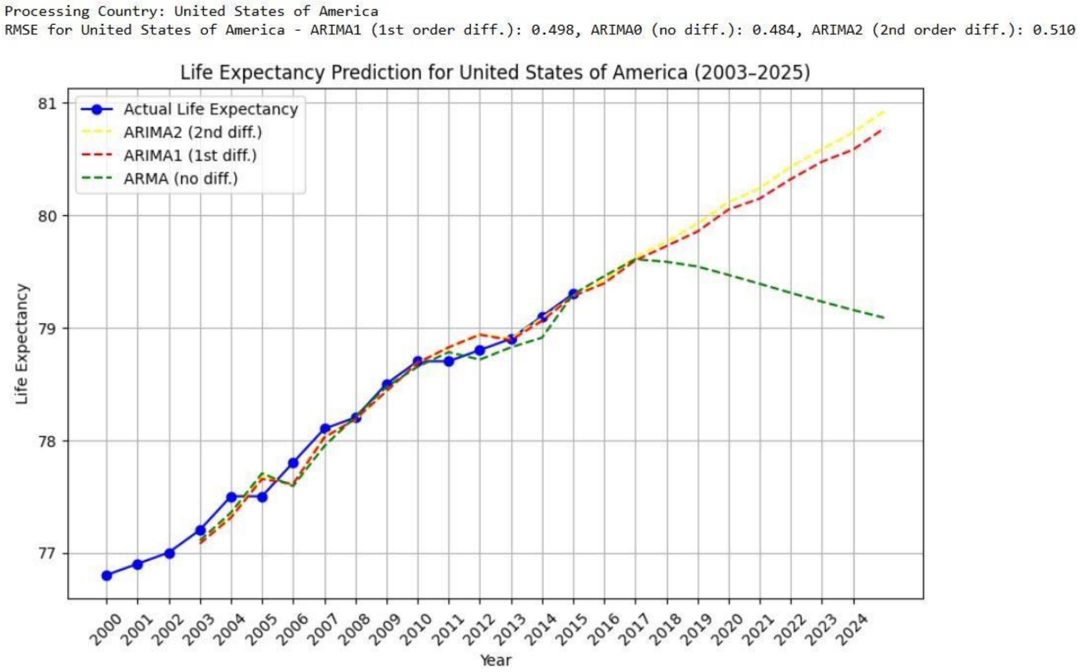
For China (figure 7), The life expectancy analysis for China shows a consistent upward trend from 2000 to 2015, increasing from approximately 72 to around 76 years.The ARIMA model without differencing closely follows the observed trend and projects a gradual increase beyond 2015, though at a slightly slower rate. It also has the lowest RMSE, indicating the best overall fit. The ARIMA model with first-order differencing aligns well with past data but shows a steeper upward trajectory in its forecasts, with a slightly higher RMSE. The second-order differencing ARIMA model captures historical variations but predicts a sharp increase in life expectancy, leading to greater forecasting variance and the highest RMSE. Overall, the ARIMA model without differencing demonstrates the best balance between accuracy and long-term trend estimation, making it the most suitable choice for life expectancy prediction in China.



```
Processing Country: China
RMSE for China - ARIMA1 (1st order diff.): 0.887, ARIMA0 (no diff.): 0.865, ARIMA2 (2nd order diff.): 0.909
```

Figure 7 Predictions of China

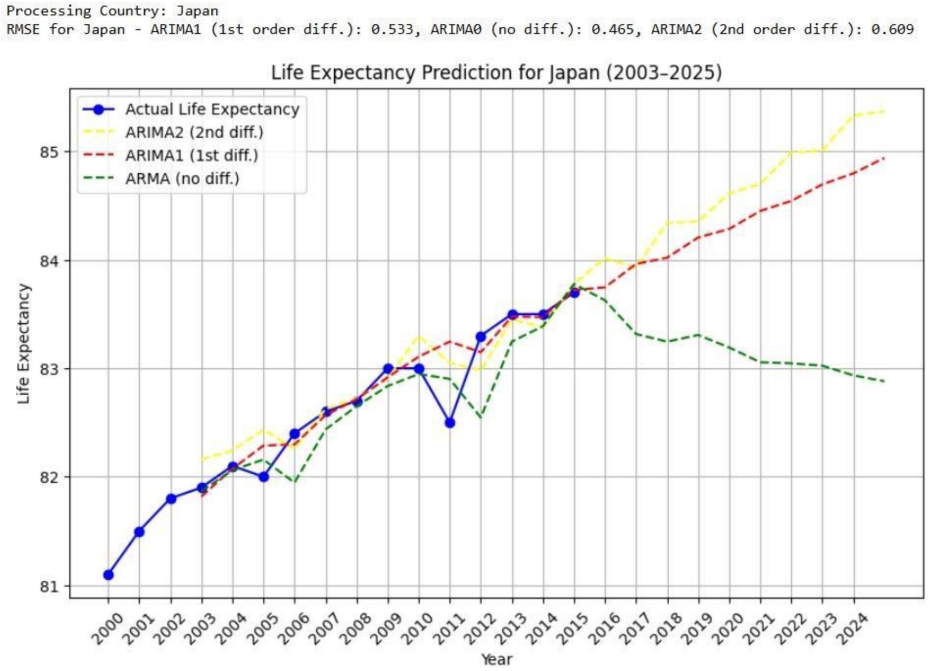
For United States of America (figure 8), the ARIMA model with no differencing (ARIMA0) has the lowest RMSE at 0.484, indicating the best fit among the tested models. The ARIMA1 model with first-order differencing follows closely with an RMSE of 0.498, while ARIMA2, which applies second-order differencing, has a slightly higher RMSE of 0.510. The ARMA model, represented by the green dashed line, diverges downward after 2017, suggesting that it may not be suitable for long-term forecasting in this case.



```
Processing Country: United States of America
RMSE for United States of America - ARIMA1 (1st order diff.): 0.498, ARIMA0 (no diff.): 0.484, ARIMA2 (2nd order diff.): 0.510
```

Figure 8: Predictions of United States of America

The life expectancy analysis for Japan (Figure 9) reveals a upward trend from 2000 to 2015, with actual life expectancy rising from approximately 81 years to around 84 years. The ARIMA model without differencing, represented by the green dashed line, aligns well with historical data but starts diverging after 2015, showing a slight decline. The ARIMA model with first-order differencing, shown in red, closely follows the observed values and projects a steady increase in life expectancy beyond 2015. The second-order differencing ARIMA model, depicted in yellow, also captures the historical trend but predicts a more pronounced upward trajectory. Among these models, ARIMA without differencing has the lowest RMSE, indicating the best overall fit for Japan's life expectancy prediction.



```
Processing Country: Japan
RMSE for Japan - ARIMA1 (1st order diff.): 0.533, ARIMA0 (no diff.): 0.465, ARIMA2 (2nd order diff.): 0.609
```

Figure 9: Predictions of Japan

MODEL	DIFFERENCING ORDER	RMSE
ARMA	d=0	1.339
ARIMA1	d=1	1.748
ARIMA2	d=2	2.754

Table1: Comparision of overall RMSE values of models.

Overall, the results suggest model ARIMA with no differencing (ARMA) perform best for forecasting life expectancy in both countries.

V. Conclusion

This study compared the forecasting potential of ARIMA with different orders of differencing in predicting life expectancy trends for various nations. By evaluating model performance using Root Mean Squared Error (RMSE), differences in accuracy were identified. The results indicate that the ARIMA model without differencing performed best. The first-order differencing model exhibited a slightly higher error, while the second-order differencing model showed greater variation. These results indicate that the ARIMA model without differencing is the most suitable for long-term life expectancy forecasting, as it ensures lower error and maintains a stable trend estimation.

References

[1] Schöley, Jonas, et al. "Life expectancy changes since COVID-19." Nature human behaviour 6.12 (2022): 1649- 1659.

[2] Woolf, Steven H., Ryan K. Masters, and Laudan Y. Aron. "Changes in life expectancy between 2019 and 2020 in the US and 21 peer countries." JAMA Network Open 5.4 (2022): e227067-e227067.

[3] Aburto, José Manuel, et al. "Dynamics of life expectancy and life span equality." Proceedings of the National Academy of Sciences 117.10 (2020): 5250-5259.

[4] Stewart, Susan T., David M. Cutler, and Allison B. Rosen. "Forecasting the effects of obesity and smoking on US life expectancy." New England Journal of Medicine 361.23 (2009): 2252-2260.

[5] Perumalsamy, Jegatheeswari, Bhargav Kumar Konidena, and Bhavani Krothapalli. "AI-Driven Risk Modeling in Life Insurance: Advanced Techniques for Mortality and Longevity Prediction." Journal of Artificial Intelligence Research and Applications 3.2 (2023): 392-422.

[6] Bali, Vikram, et al. "Life Expectancy: Prediction & Analysis using ML." 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). IEEE, 2021.

[7] Ye, Wenjing, et al. "Identification of COVID-19 clinical phenotypes by principal component analysis-based cluster analysis." Frontiers in medicine 7 (2020): 570614.

[8] Choubey, Dilip K., et al. "Comparative analysis of classification methods with PCA and LDA for diabetes." Current diabetes reviews 16.8 (2020): 833-850.

[9] Margaritis, Alexandros, et al. "Identification of ageing state clusters of reclaimed asphalt binders using principal component analysis (PCA) and hierarchical cluster analysis (HCA) based on chemo-rheological parameters." Construction and Building Materials 244 (2020): 118276.

[10] Gadekallu, Thippa Reddy, et al. "Early detection of diabetic retinopathy using PCA-firefly based deep learning model." Electronics 9.2 (2020): 274.

[11] He, Yunting, et al. "Moving average based index for judging the peak of the COVID-19 epidemic." International Journal of Environmental Research and Public Health 17.15 (2020): 5288.

[12] Singh, Ram Kumar, et al. "Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model." JMIR public health and surveillance 6.2 (2020): e19115.

[13] Su, Yaxi, Chaoran Cui, and Hao Qu. "Self-attentive moving average for time series prediction." Applied Sciences 12.7 (2022): 3602.

[14] Perone, Gaetano. "An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy." MedRxiv (2020): 2020-04.

[15] Ospina, Raydonal, et al. "An overview of forecast analysis with ARIMA models during the COVID-19 pandemic: Methodology and case study in Brazil." Mathematics 11.14 (2023): 3069.

[16] Tran, Thai T., L. T. Pham, and Q. X. Ngo. "Forecasting epidemic spread of SARS-CoV-2 using ARIMA model (Case study: Iran)." Global Journal of Environmental Science and Management 6.Special Issue (Covid-19) (2020): 1-10.

[17] Dehesh, Tania, Heydar Ali Mardani-Fard, and Paria Dehesh. "Forecasting of covid-19 confirmed cases in different countries with arima models." MedRxiv (2020): 2020-03.

[18] Li, Xiao, et al. "Research on the prediction of dangerous goods accidents during highway transportation based on the ARMA model." Journal of loss prevention in the process industries 72 (2021): 104583.

[19] Awariefe, C., and G. Ekruyota. "MODELING AND FORECASTING LIFE EXPECTANCY AT AGE SIXTY- FIVE IN THE POPULATION OF NIGERIA USING AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODELS." The Journals of the Nigerian Association of Mathematical Physics 65 (2023): 87- 94.

[20] Ayele, Amare Wubishet, Mulugeta Aklilu Zewdie, and Tizazu Bayko. "Modeling and forecasting the global daily incidence of novel coronavirus disease (COVID-19): An application of autoregressive moving average (ARMA) model." International Journal of Public Health and Safety 5 (2020).