Forecasting of Global Life expectancy using SARIMA models

Venkata Krishna Rao Maddela
Dept. of Computer Science & Engineering (Data Science), Institute of Aeronautical Engineering, Hyderabad, India

Abstract: Forecasting of life expectancy of a country enables healthcare industry and government to plan and execute the future health programs by appropriate resource allocation, long-term strategy and assess the demands on healthcare into the future. In this study a univariate seasonal autoregressive integrated moving average model is used to model the historical data of annual life expectancy of individual countries and groups of countries, 266 in total. Stationary tests are used to estimate nonstationarity. Difference tests are used to estimate the integrated order of the timeseries. The seasonal periodicities are estimated from Fourier transform of the data. Best model parameters are estimated using Broyden Fletcher Goldfarb Shanno (BFGS) algorithm. The algorithm searches for using grid search over the parameter space. The selected parameters are fine-tuned using 5-fold cross validation. The best model is fitted to test data and error metrics are computed. Finally, the Life expectancy is forecasted along with confidence intervals for the next 10 years into the future for each country. Python programming framework is used in the study.

Keywords: SARIMA model, ARMA model, Pole-Zero diagram, Nonstationary tests, Seasonal period estimation Timeseries forecasting, Timeseries Integrated Order

1. Introduction

Life Expectancy is one major indicator of citizen's quality life in a country or a region. Forecasting life expectancy of a country enables healthcare industry and government to plan and execute the future health programs by appropriate resource allocation, long-term strategy and assess the demands on healthcare into the future. The forecasting of life expectancy can be done by using traditional timeseries models such as seasonal autoregressive integrated moving average (SARIMA) model [1] and its variants [2, 3]. The (S)AR(I)MA model was successfully applied in the analysis and prediction in applications such as seismology, [4], biomedical signals [5], traffic-signal control [6], communications [7], radar systems [8] and many more. Speech data form a quasi-stationary timeseries and extensively studied in the name of linear prediction analysis to identify the time-varying formants, frequencies at which vocal cavity resonates [9, 10]. The time evolution of phonetic-acoustics events in time-frequency plane was studied in an Indian language [11, 12, 13]. Statistical characteristics (moments and joint moments) of a nonstationary sequence vary with time, making it very difficult to model, predict and forecast the timeseries accurately. Several statistical tests were developed to detect nonstationarity in the timeseries. Some popular tests used to identify the nonstationarity in a timeseries [1, 2, 3] are tabulated in Tables 1 and 2. A nonstationary timeseries can be converted to a stationary timeseries by differencing the given series d-times where d is a small integer. In most cases $d = 1 \; or \; 2$ must be sufficient; $d = 3$ being a rare case, and more than 3 may not be required at all, in real world scenarios. The difference tests are designed to estimate the n value. The *nsdiff()* test is used for seasonal differencing, while *ndiff()* test is used for nonseasonal differencing. In fact *ndiff()* uses one of the ADF, KPSS and PP tests; and *nsdiff()* uses CH Test and OCSB Test. Information criteria such as Akaike information criterion (AIC) are used to compute d-value.

Table 1
Nonstationarity Tests

Test	Null Hypothesis ($H0$)	Alternate Hypothesis ($H1$)
Augmented Dickey-Fuller (ADF)	Timeseries are nonstationary i.e. has a unit root	Timeseries is stationary
Kwiatkowski-Phillips-Schmidt-Shin (KPSS)	Timeseries are trend stationary	Timeseries is nonstationary
Phillips-Perron (PP)	Timeseries is nonstationary	Timeseries is nonstationary

Table 2

Difference Tests

Test	Purpose of the test
Osborn-Chui-Smith-Birchenhall (OCSB)	To find number of times, the difference operation is to be applied for converting to stationary series
Canova-Hansen (CH)	To find number of times, the difference operation is to be applied for converting to stationary series

In present study, a univariate seasonal autoregressive integrated moving average (SARIMA) model is used to model the historical data of annual life expectancy of individual countries and groups of countries. The model parameters are selected using the grid search over the allowable parameter space, using BFGS algorithm while minimizing the corrected AIC. Using the best model the life expectancy for the years (2014 – 2023) and error metrics are computed. The life expectancy for the period 2024 – 2033 is forecasted. The predictions and forecasts are computed for all countries, but results are presented for 18 countries, which are selected based on their unique trends/patterns in the historical data. The confidence intervals for the forecasts are computed.

Rest of the paper is framed as follows. Section II reviews the related work on the timeseries forecasting using ARIMA and related models along with applications of timeseries models recently used. Section III presents the proposed methodology and describes the dataset used in the study. Section IV presents the modelling experiments and analysis results of the proposed methodology. In section V, conclusions on the present work are drawn and future scope of work is projected.

2. Literature Review

This section provides a review of related work chronologically year-wise. In [**14**] the demand in a food company was forecasted using ARIMA model by using Box–Jenkins procedure. The model was evaluated using four criteria: Akaike, Schwarz Bayesian, standard error and maximum likelihood. The resultant forecasts on food demand were used to make the appropriate production planning to eliminate food waste and avoid cost losses. Hosam H. A. et. al., [**15**] forecasted the water consumption and revenues from water service using Auto-Regressive Integrated Moving Average (ARIMA), Hybrid ARIMA, Singular Spectrum Analysis (SSA), and Linear Regression. A dataset collected from Khan Younis municipality (KHM) was used in the study and identified ARIMA-ANN as the best algorithm. The maximum water consumption five years after 2017 would increase by about 8.4%. In [**16**] linear ARIMA and Holt's model were studied among other non-linear methods like neural network auto-regressive model. The system was aimed at arriving the best model to predict the risk of deaths and infections. The second task was the implementation of the third wave of infections and deaths in Russia. The study prompted about the limitations of study during time-changing conditions, and state that time-series forecasting can be accurate only in the short term and suggests for nonlinear timeseries models. The results have shown the importance of seasonality in timeseries modelling. However, estimation of seasonal period is still a challenge and in general grid search is used. In [**17**] a study was made to examine the impact of government's policy intervention on dispensing of the medicine, quetiapine. The dispensing claims data was used to build an ARIMA model A study was made to examine the impact of government's policy intervention on dispensing of quetiapine. The dispensing claims data was used to build an ARIMA model using *auto.arima()* method of the forecast package in R. The study highlighted the advantages of ARIMA model compared to segmented regression (SR) model. A study on Covid-19 cases in India [**18**] utilized a hybrid model comprising ARIMA model and autoregressive neural network for extracting linear and nonlinear relationships in the data respectively. The study claims that low error metrics were achieved by the hybrid model compared to ARIMA model and even LSTM model for short-term prediction. In [**19**] the aircraft failure rate was analysed by Grey Verhulst Model, ARIMA and Artificial Neural Network with back propagation. These models were evaluated using the traditional error measures: MAE, RMSE and PMAE, and other metrics: Nash-Sutcliffe Efficiency coefficient (NSE), Equal Coefficient (EC), Index of Agreement (IA) and Theil Inequality Coefficient (TIC). The best ARIMA model was used to predict and forecast aircraft failure rates. The results demonstrated that the combination model performed much better than any of the single models. Also the combination model based on Induced ordered weighted averaging technique worked better than other combination models.

Kin Wai Ng. et. al. [**20**] analyzed the social media activity on different topics from geo-political contexts such as Venezuelan Political Crisis of 2019, Chinese-Pakistan Economic Corridor, and Belt and Road Initiative in East Africa using three models: ARIMA, Hawkes processes and Shifted baseline. Using the best fit model social media activity was forecasted for a week into the future at granularity of one hour. The results showed that ARIMA model performed the best among the three techniques, which highlighted the capability of ARIMA model to flexibly model the trends, autocorrelation and seasonality of different types of impacts. In [**21**] ARIMA model was applied on telecommunication subscriber usage data to predict the growth in the data usage. In [**22**] author did several experiments on using *auto.arima()* function and found that the usage auto.*arima()* is very tricky. One must carefully see the optional hyperparameters available for *auto.arima()*, especially parameter search method: step search, random search, or grid search. In case of first two methods, there is a possibility that the best model may be missed. In [**23**] an ARMA and ARIMA models were fitted to annual mean temperature data. The ARMA fitted better than ARIMA, which brings out the importance of stationarity tests in timeseries analysis.

Univariate and vector ARMA models, and dynamic linear models were reported on epilepsy research using R-environment in [**24**]. The review study of [**25**] broadly includes linear and nonlinear timeseries methods in theory. Some analytical expressions for ARMA and GARCH models were presented. However, no simulations or experiments on real data were conducted to support the theory. US healthcare expenditure was forecasted using both Random Forest and ARIMA models with unity integrated order in [**26**]. The study highlights that the random forest of machine learning could not perform well compared to ARIMA model. In [**27**] an over-identified system comprising of the features like government health per capita expenditure was developed and the correlations among the features was studied. The study has shown objectively that the government's expenditure on public health significantly improves citizen's average health status. In a study [**28**] ARIMA model was applied to analyze the five timeseries: gross domestic product, unemployment rate extended National Consumer Price, and so on. Five different models, one for each timeseries, was trained on the historic data using optimum model orders. The best fit models were used to forecast the series next 5 years into the future. In [**29**] monthly water level data of Morava e Binçës river from 2014 to 2021 was analysed using ARIMA and Error Trend and Seasonality (ETS), or Exponential Smoothing (ES) models. Best fit model was used to predict the periods during which low and high-water levels may occur between 2022 and 2024 were predicted. Saleh Al Sulaie., [**30**] applied an ARIMA model to analyse the number of traffic crashes in Saudi Arabia for the period: 2002-2022. The number of injuries per each 1,000 traffic accidents decreased, even though there was an increasing trend in the number actual accidents. The number of injuries was forecasted using the best fitted ARIMA model.

In [**31**] regression in combination with ARIMA was used to fit timeseries model to the stock prices. The combination was used to counter the limitations of ARIMA in forecasting long into the future. Here the regression captures seasonal patterns effects. Though the conceptual reason was not explicitly stated in that reference, author conjectures that if the residual after regression model is not purely random and correlated, then fitting ARMA model to the residual decorrelates the residual. In [**32**] Long Short-Term Memory networks and ARIMA were used to model the stock prices of Apple Inc., for the years 2016 to 2024. The residual from the ARIMA model was fed to LSTM model appropriately gated. The metrics RMSE, MAE, Directional accuracy and R^2 value was found to be the better for the hybrid model. The study of [**33**] analyzed the microbial density for monitoring water purification system. The data was first aggregated cumulative logarithmically transformed and then the linear, exponential and Holt-Winters, ARIMA methods were fitted. In [**34**], the monthly London Metal Exchange price of base metals was forecasted using an autoregressive Light Gradient Boosting Machine (LightGBM) and by a combination of the LightGBM algorithm and ARIMA model. The performance of these two models was compared with performance of three benchmark models: a global mean (GMM) model, an ES model and an ARIMA model. The LightGBM model outperformed the benchmark models in forecasting the price for the next 6 months into the future for nickel and aluminium returns, However, the ensemble method performed well in case of zinc and copper returns. In case of forecasting of tin and lead returns, the ARIMA model performed well. In a very recent review study [**35**] on SARIMA models in python environment, the model issued such as stability, nonstationarity and invertibility were discussed in detail using pole-zero diagrams of the model. The study of [**36**] explores the ARIMA novel to predict CPC pressure based on head losses. The study used the Box-Jenkins methodology to analyze timeseries and the proposed system enhances the identification of anomalies in pressure measurements and thus enabling efficient pressure management in the network. In a very recent work [**37**], a SARIMA model was applied on WHO's health expenditure data of 183 countries. The seasonal period was estimated using Fourier transform, and the health expenditures of all countries were forecasted for the period 2025-29.

In the present study, SARIMA model is used to fit the global life expectancies. In the next section, the theoretical background and algorithms of the proposed models are presented.

2.1. Theoretical framework

A general time series model $SARIMA\ (p,d,q)(P,D,Q)_m$ is mathematically specified [2, 3] as recursive equation given below

$$y_n = (\varphi_1 y_{n-1} + \varphi_2 y_{n-2}, \dots, +\varphi_p y_{n-p}) + (\varphi_{1s} y_{n-m} + \varphi_{2s} y_{n-2m}, \dots, +\varphi_{Ps} y_{n-Pm}) \\ +e_n + (\theta_1 x_{n-1} + \theta_2 x_{n-2}, \dots, +\theta_q x_{n-q}) + (\theta_{1s} x_{n-m} + \theta_{2s} x_{n-2m}, \dots, +\theta_{Qs} x_{n-Qm}) \quad \dots\dots\dots \quad (1)$$

where:

- $\boldsymbol{y}$ is the time series, $\boldsymbol{x}$ is the driving force,
- *(p, q)* are the orders of AR and MA parts (nonseasonal),
- *(P, Q)* are the orders of AR and MA parts (seasonal),
- *(d, D)* are nonseasonal and seasonal integrated orders,
- B is the unit-delay operator,
- $\varphi_1, \varphi_2, \dots, \varphi_p$ are the parameters of nonseasonal AR part,
- $\theta_1, \theta_2, \dots, \theta_q$ are the parameters of nonseasonal MA part,
- $\varphi_{1s}, \varphi_{2s}, \dots, \varphi_{Ps}$ are the parameters of seasonal AR part
- $\theta_{1s}, \theta_{2s}, \dots, \theta_{Qs}$ are the parameters of seasonal MA part.

Autocorrelation and partial correlations are useful exploratory tools of a timeseries. The autocorrelation of a stochastic timeseries $\boldsymbol{y}$ having statistical mean μ_x and variance σ_x^2 is given by [6]

$$\rho_k = E\big[(y_n - \mu_y)(y_{n+k} - \mu_y)\big]/\sigma_y^2 \quad \dots\dots\dots \quad (2)$$

and the one-sided sample autocorrelation estimate of a finite timeseries of length *N* is given by

$$\hat{\rho}_k = \sum_{n=0}^{N-1} (y_n - \bar{y})(y_{n+k} - \bar{y})/(y_n - \bar{y})^2 \quad k = 0, 1, 2, \dots, N-1 \quad \dots\dots\dots \quad (3)$$

where $\bar{y}$ is the sample mean of the finite timeseries. The partial autocorrelation of the timeseries is given by

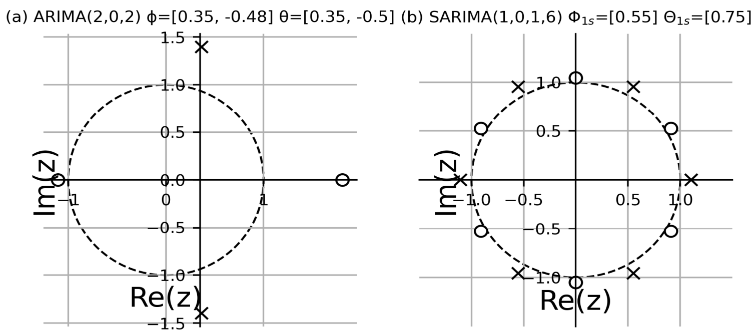
$$\varphi_{n,n} = \big(\rho_n - \sum_{k=1}^{N-1} \varphi_{n-1,k}\rho_{n-k}\big)/\big(1 - \sum_{k=1}^{N-1} \varphi_{n-1,k}\rho_k\big) \quad \dots\dots\dots \quad (4)$$

and its estimate $\widehat{\varphi}_{n,n}$ is obtained by replacing the parameters $\varphi_{n-1,k}$ by their estimates $\widehat{\varphi}_{n-1,k}$ and replacing the correlations ρ_{n-k} and ρ_k respectively by their estimates $\hat{\rho}_{n-k}$ and $\hat{\rho}_k$ in Eq (4). If longer tails with gradual decay are observed in the (partial) autocorrelation, it signals a nonstationarity in the timeseries. Statistical characteristics (moments and joint moments) of a nonstationary sequence vary with time, making it very difficult to model, predict and forecast the timeseries accurately. The system equation corresponding to nonseasonal part of the recursive equation Eq (1) is given by [7]

$$\mathrm{H}(z) = \frac{\Theta(z)}{\emptyset(z)} = \frac{1 + \theta_1 z + \theta_2 z^2, \dots, +\theta_q z^q}{1 + \varphi_1 z + \varphi_2 z B^2, \dots, +\varphi_p z^p} \quad \dots\dots\dots (5)$$

where $\emptyset(z)$ is nonseasonal AR polynomial, $\Theta(z)$ is nonseasonal MA polynomial, $\emptyset_s(z)$ is seasonal AR polynomial and $\Theta_s(z)$ is seasonal MA polynomial. For the timeseries to be stationary, the polynomials $\emptyset(B)$ and $\emptyset_s(B)$ must have roots outside the unit circle in z-plane. For the model to be invertible, the polynomials $\Theta(B)$ and $\Theta_s(B)$ must have roots outside the unit circle. Invertibility means the model can be inverted to derive the driving noise $\boldsymbol{x}$ from the timeseries $\boldsymbol{y}$, so that estimate of driving noise variance $\hat{\sigma}_x^2$ (which is also one of the model parameters) can be obtained. Thus, random selection of MA and AR coefficients may not result in a causal (realizable), stable and invertible system/model. A timeseries is simulated using the parameters: $[\varphi_1, \varphi_2] = [0.35, -0.48], \mathrm{p} = 2, \varphi_{1s} = 0.55, P = 1, [\theta_1, \theta_2] = [0.35, -0.5], \mathrm{q} = 2, \theta_{1s} = 0.75, Q = 1$. The poles and zeros of the simulated model are computed by using the built-in method *.roots()* of *numpy* package, is visualized using python code (customized) in Fig 1. In the figure, the marker 'o' represents zero and the marker 'x' represent a pole. The roots of nonseasonal AR (poles) and MA (zeros) polynomials are shown in Fig 1(a). The roots of seasonal AR (poles) and MA (zeros) polynomials are shown in Fig 3(b) The simulated system is stable and invertible, as all the roots are outside the unit circle.

Figure. 1.
Pole-Zero diagram of an example ARIMA timeseries (a). Nonseasonal model (b). Seasonal model.



3. Research Methodology

Based on theoretical information provided in section 2.1, the research methodology is proposed, and the algorithm is developed in this section. The dataset details are also provided in this section.

3.1. Participants

In present study, the secondary dataset of global annual life expectancy available at WHO website [**50**] is used. The dataset gives the annual life expectancy of 266 individual countries or groups of countries in total for the years 1960 to 2023. Some groups of countries specified in the dataset are International Development Association (IDA) countries, International Bank for Reconstruction and Development (IBRD) countries, heavily indebted poor countries (HIPC) countries, IBRD only, IDA & IBRD total, IDA total, IDA blend, IDA only and so on. The data of all 266 samples (rows of .CSV data file) are analyzed using the proposed methodology. The results of analysis are presented for only 18 countries, which are selected based on their unique trends/patterns in the data.

3.2. Instruments

An 70% (i.e. for the period 1960 – 2003) of timeseries is used for training an ARIMA model. The model parameters are selected using the grid search over the allowable parameter space, using BFGS algorithm while minimizing the corrected AIC. The trained model is validated using 5-fold cross validation by adding an additional 15% of data to the training data. Using the best model the life expectancy for the years 2014 – 2023 (test data) is predicted and error metrics are computed. The life expectancy for the period 2024 – 2033 is forecasted. The predictions and forecasts are computed for all countries, but results are presented for 18 countries, which are selected based on their unique trends/patterns in the historical data. The trend or profile of the forecast appears to be in line with that of the historical data subjectively. No objective measures are computed as there is no ground truth for forecasts. However, the confidence intervals for the forecasts are computed.

Algorithm for ARIMA Model building and forecasting

Step 1. Initiate the country serial no (n) =1.

Step 2. Read the life expectancy timeseries data for the years 1960 to 2024 of the country(n). Check for any missing data for any year. If data is missing, go to Step 19.

Step 3. Segment the timeseries data into two parts: training data (first 70% of samples) and test data (last 30% of samples).

Step 4. Conduct the stationarity tests: ADF and KPSS tests and identify the training data as stationary (S) or nonstationary (NS).

Step 5. Conduct the difference tests and estimate number of nonseasonal successive differences $\hat{\boldsymbol{d}} = (\hat{d}_1, \hat{d}_2, \dots \dots)$ required to convert the training data to stationary data.

Step 6. Conduct the seasonal difference tests and estimate number of seasonal successive differences ($\hat{\boldsymbol{D}} = (\hat{D}_1, \hat{D}_2, \dots \dots)$) required to convert the training data to a stationary one.

Step 7. Apply successive differences as specified by $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{D}}$.

Step 8. Compute full autocorrelation (ACF) and partial autocorrelation (PACF) of the training data using the expressions (2) and (3).

Step 9. Visualize the correlation patterns optionally.

Step 10. Apply a threshold on ACF and PACF and identify the lags ($\hat{m}_1, \hat{m}_2, \dots \dots$) with correlation values above the threshold. These are the seasonal periods estimated from the correlation analysis of the timeseries.

Step 11. Compute Fourier spectrum of the training data and identify peaks and their frequencies $\hat{f}_1, \hat{f}_2, \dots$). in the spectrum. These are the seasonal periods estimated from the frequency analysis of the timeseries.

Step 12. Compute the seasonal periods $(1/\hat{f}_1, 1/\hat{f}_2, \dots)$.

Step 13. Create the combined vector of the estimated seasonal periods: $\hat{\boldsymbol{m}} = (\hat{m}_1, \hat{m}_2, \dots, 1/\hat{f}_1, 1/\hat{f}_2, \dots)$ using the seasonal periods estimated from ACF and Fourier transform.

Step 14. Compute the maximum likelihood Estimate of the model orders $\hat{p}, \hat{q}, \hat{P}, \hat{Q}$ by applying Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm of the training data.

Step 15. Repeat step 14 for all parameters (grid search) over the parameter space: $(1, 2, \dots, p_{max})$, $(1, 2, \dots, q_{max})$, $(1, 2, \dots, P_{max})$, $(1, 2, \dots, Q_{max})$ where p_{max}, q_{max}, P_{max} and Q_{max} are predefined maximum orders.

Step 16. Select the optimum model parameters p_{opt}, q_{opt}, P_{opt} and Q_{opt} that give the minimum value of Akaike Information Criterion (AIC).

Step 17. Using the optimum model, predict the timeseries for the duration of testing data.

Step 18. Compute the error metrics: MAE, MAPE and RMSE using the expressions: $\text{MAE} = \sum_{k=1}^{M} |y_k - \hat{y}_k|$ $\text{MAPE} = (1/M) \sum_{k=1}^{M} |y_k - \hat{y}_k| / |y_k| \times 100$ and $\text{RMSE} = \sqrt{(1/M) \sum_{k=1}^{M} (y_k - \hat{y}_k)^2}$.

Step 19. Forecast the life expectancy for the next 6 years into the future.

Step 20. Plot the full time series data as dashed line along with markers. Overlay the plots of the predicted data and the forecasted data with a different color. Overlay the confidence intervals for the forecasts as patch area.

Step 21. Increment the country's serial number**:** $n = n + 1$. If $n < N_c$ (total number of countries), go to Step 2.

Step 22. Stop

4. Implementation and Results

This work is implemented using python programming language in the framework of Anaconda distributed spider IDE on the laptop. The range of time variable is 1960 – 2023, with a time resolution of one year. The life expectancy of 18 countries: Aruba, United Arab Emirates, Argentina, Burkina Faso, Bermuda, China, Cote dIvoire, Denmark, Estonia, France, Micronesia Fed. Sts., India, Japan, Madagascar, Northern Mariana Islands, Solomon Islands, United States and South Africa are shown in Fig. 2 and Fig. 3. The countries are selected so that distinct trend patterns are included in the study. The ACF and PACF of these timeseries are shown in Fig. 4 through Fig. 7 respectively. The full autocorrelations have long tails indicating nonstationarity. The confidence interval is shown in pale blue, at first lag for all countries. However, there are few significant lags within the confidence interval touching the boundary.

Figure. 2.
Life expectancy plots of 1 - 9 countries (1960 to 2023)

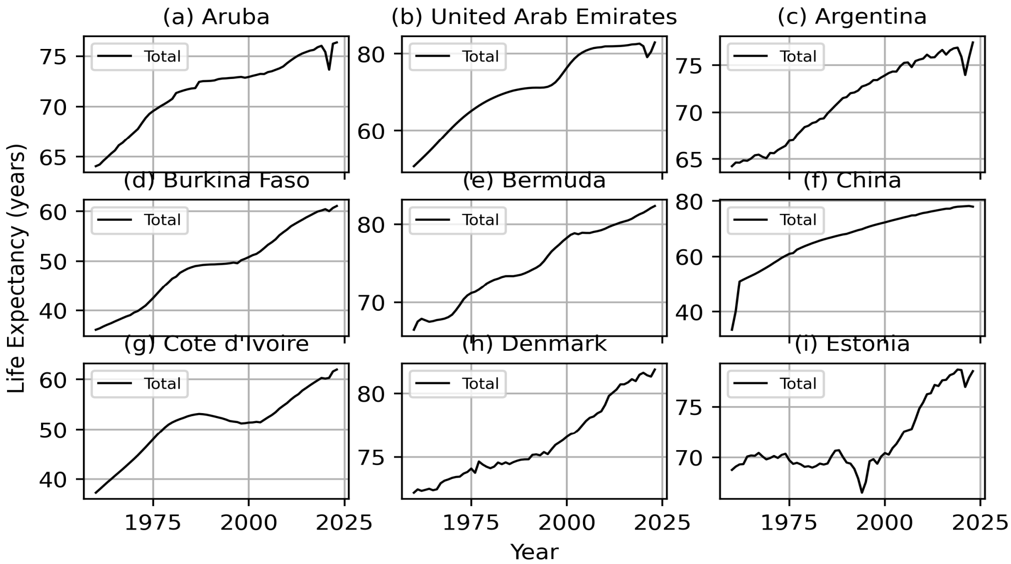


Figure. 3.
Life expectancy plots of 10 - 18 countries (1960 to 2023)

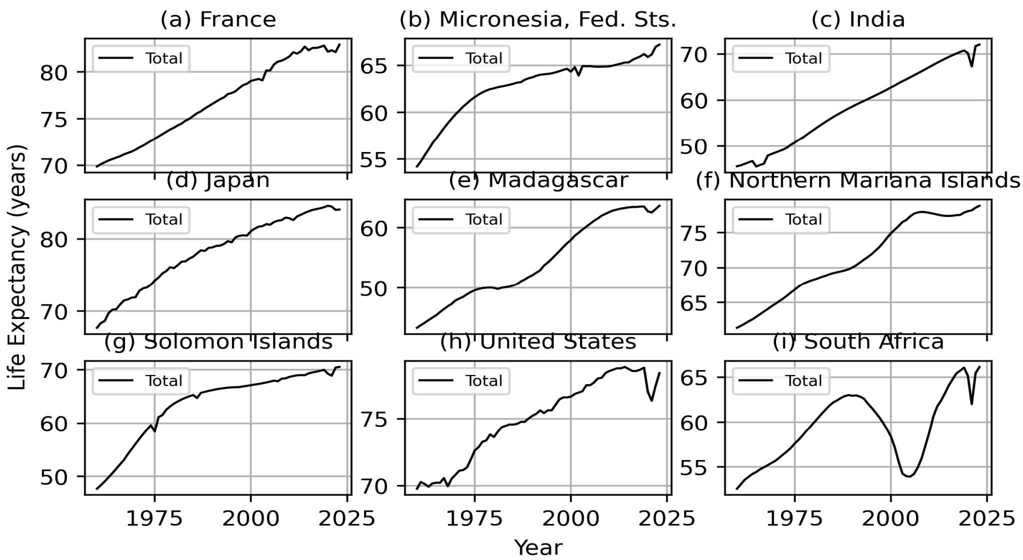


Figure. 4.
Autocorrelations of Life expectancies of first set of 9 countries

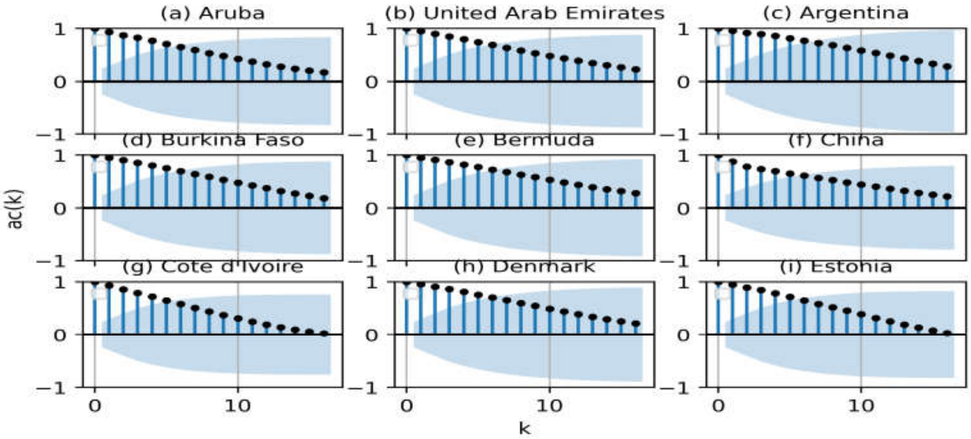


Figure. 5.
Autocorrelations of Life expectancies of second set of 9 countries

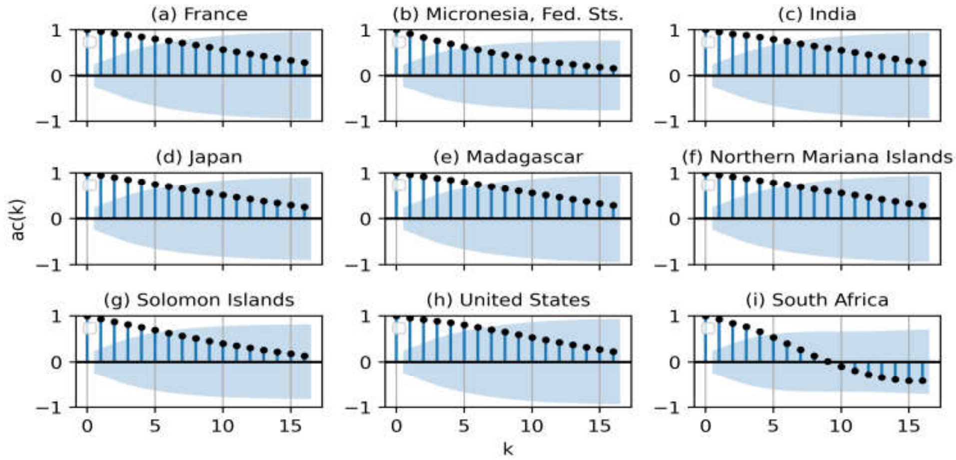


Figure. 6.
Partial autocorrelations of Life expectancies of first set of 9 countries

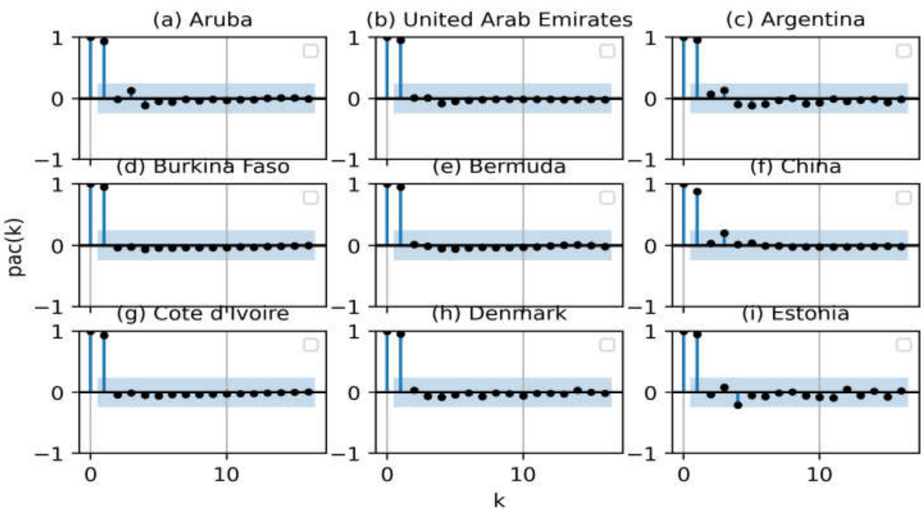
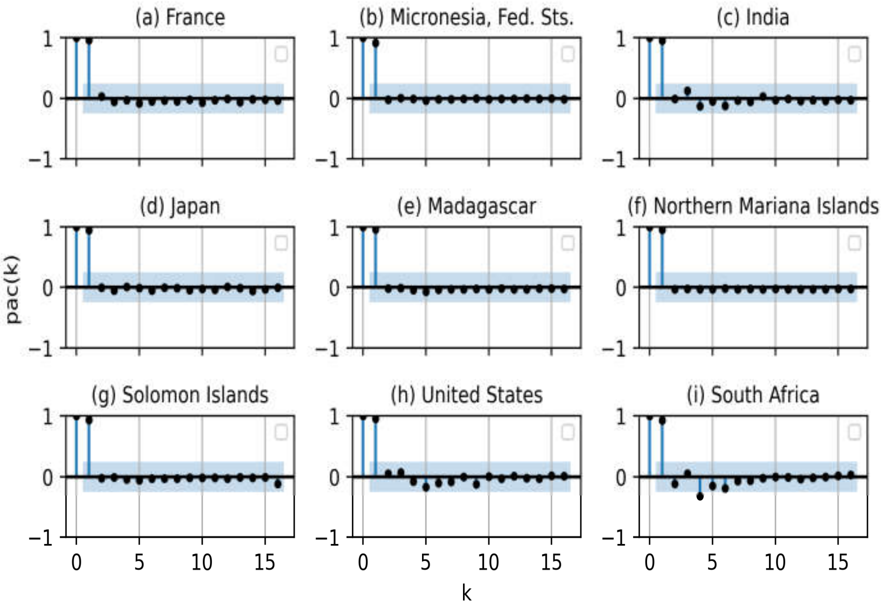


Figure. 7.
Partial autocorrelations of Life expectancies of second set of 9 countries



The fourier transforms of the life expectancy of 18 countries are shown in Fig. 8 and Fig. 9. The peaks in spectrum are shown as red dots. Each peak indicates a periodicity in the timeseries and their location is the frequency. The reciprocal of the frequency gives the periodicities in the timederies. The first peak is to be ignored as it indicates the period ($m = 1$) corresponding to the sampling index of the original timeseries. The fouries coefficient at zero frequency is always zero, since mean is subtracted from th data. The stationarity tests (ADF and KPSS) and difference tests are conducted on the data. The results of stationarity tests are shown in Table 3 and 4 respectively. The columns are the t-Stat (test statistic), pValue (probability value), cValue (critical value for the t-Stat at 5%), number of lags used (#Lags), and the number of observations used (#Obs). If pValue > α, the significant level (here it is 0.05) the null hypothesis H_0 is rejected, and the alternative hypothesis H_1 is accepted. It may be noted that the null hypothesis and alternative hypothesis are different for different tests as given in Table 1 and Table 2. The decision of nonstationarity is given in the last column (NS for nonstationary and S for stationary). The results of difference tests are shown in Table 12. The column 0 is the row index. Columns 1-3 are the number of nonseasonal differences (d) to convert the timeseries into stationary one. Column 4 is the maximum of the columns 1-3. Columns: 5-6 are the number of seasonal differences (D) to convert the series into stationary one. The last column is the minimum of the columns 5 and 6.

Figure 8
Fourier transform of Life expectancies of first set of 9 countries (Peaks in the spectrum are shown as red markers).

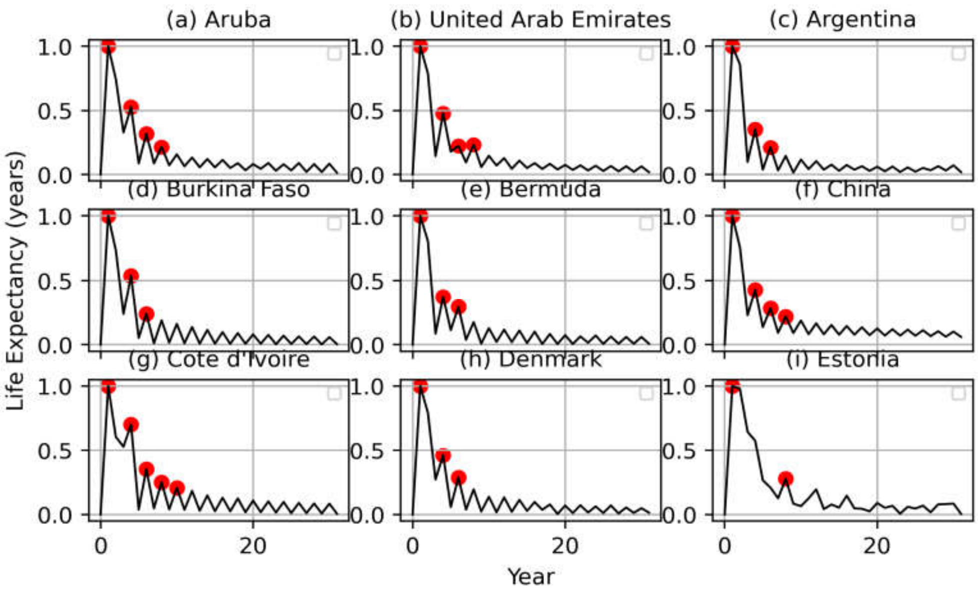


Figure 9
Fourier transform of Life expectancies of first set of 9 countries. Peaks in the spectrum are shows as red dots.

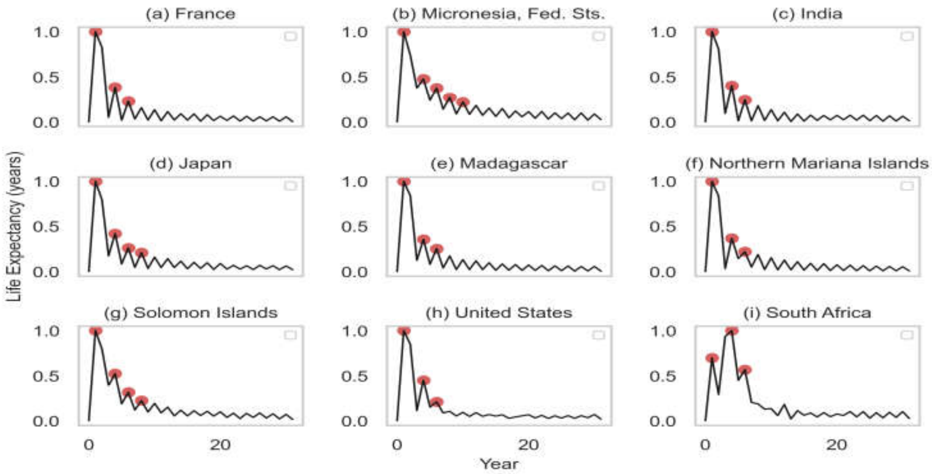


Table 3
Results of ADF test. (NS = nonstationary and S = stationary

Country	t-Stat	pValue	cValue['5%']	#Obs	#Lags	Stationary
Aruba	-2.982	0.304	-3.954	12	51	NS
United Arab Emirates	-5.089	0.001	-3.965	16	47	S
Argentina	-2.057	0.794	-3.959	14	49	NS
Burkina Faso	-2.510	0.563	-3.949	10	53	NS
Bermuda	-4.102	0.023	-3.939	5	58	S
China	-3.476	0.119	-3.940	6	57	NS
Cote dIvoire	-3.377	0.147	-3.939	5	58	NS
Denmark	-3.814	0.052	-3.965	16	47	NS
Estonia	-2.254	0.703	-3.956	13	50	NS
France	-0.197	0.998	-3.965	16	47	NS
Micronesia, Fed. Sts.	-0.414	0.997	-3.931	1	62	NS
India	-4.467	0.007	-3.937	4	59	S
Japan	-4.242	0.015	-3.930	0	63	S
Madagascar	-3.989	0.033	-3.939	5	58	S
Northern Mariana	-4.230	0.016	-3.943	7	56	S
Solomon Islands	-0.238	0.998	-3.931	1	62	NS
United States	-2.748	0.427	-3.945	8	55	NS
South Africa	-2.368	0.643	-3.956	13	50	NS

Table 4
Results of KPSS test. (NS = nonstationary and S = stationary

Country	t-Stat	pValue	cValue['5%']	#Lags	Stationary
Aruba	0.290	0.010	0.146	4	S
United Arab Emirates	0.200	0.016	0.146	4	S
Argentina	0.250	0.010	0.146	4	S
Burkina Faso	0.119	0.099	0.146	5	NS
Bermuda	0.136	0.069	0.146	4	NS
China	0.293	0.010	0.146	4	S
Cote dIvoire	0.169	0.031	0.146	5	S
Denmark	0.270	0.010	0.146	5	S
Estonia	0.261	0.010	0.146	5	S
France	0.146	0.050	0.146	4	S
Micronesia, Fed. Sts.	0.287	0.010	0.146	4	S
India	0.183	0.022	0.146	4	S
Japan	0.343	0.010	0.146	4	S

Country	t-Stat	pValue	cValue['5%']	#Lags	Stationary
Madagascar	0.118	0.100	0.146	5	NS
Northern Mariana	0.125	0.090	0.146	5	NS
Solomon Islands	0.274	0.010	0.146	5	S
United States	0.230	0.010	0.146	4	S
South Africa	0.123	0.093	0.146	5	NS

Table 5
Results of nonseasonal and seasonal difference tests

Country	ADF	PPSS	PP	*est_d*	OCSD	CH	*est_D*
Aruba	2	1	1	2	1	0	0
United Arab Emirates	2	2	1	2	3	0	0
Argentina	1	1	1	1	0	0	0
Burkina Faso	0	1	1	1	0	0	0
Bermuda	2	1	1	2	0	0	0
China	1	2	0	2	0	0	0
Cote dIvoire	0	1	2	2	1	0	0
Denmark	2	2	1	2	0	0	0
Estonia	2	1	1	2	0	0	0
France	2	1	1	2	0	0	0
Micronesia, Fed. Sts.	2	2	1	2	0	0	0
India	1	1	0	1	1	0	0
Japan	1	2	1	2	0	0	0
Madagascar	2	1	2	2	0	0	0
Northern Mariana	0	1	2	2	0	0	0
Solomon Islands	2	2	1	2	0	0	0
United States	1	1	1	1	0	0	0
South Africa	0	1	1	1	2	0	0

The forecasts of life expectancy for the next 10 years, i.e., for the years 2024-2033 for the 18 countries, are shown in Fig. 10 through Fig. 14.

Figure 10
The forecasts of health expenditures using optimum model (countries: 1 to 4)

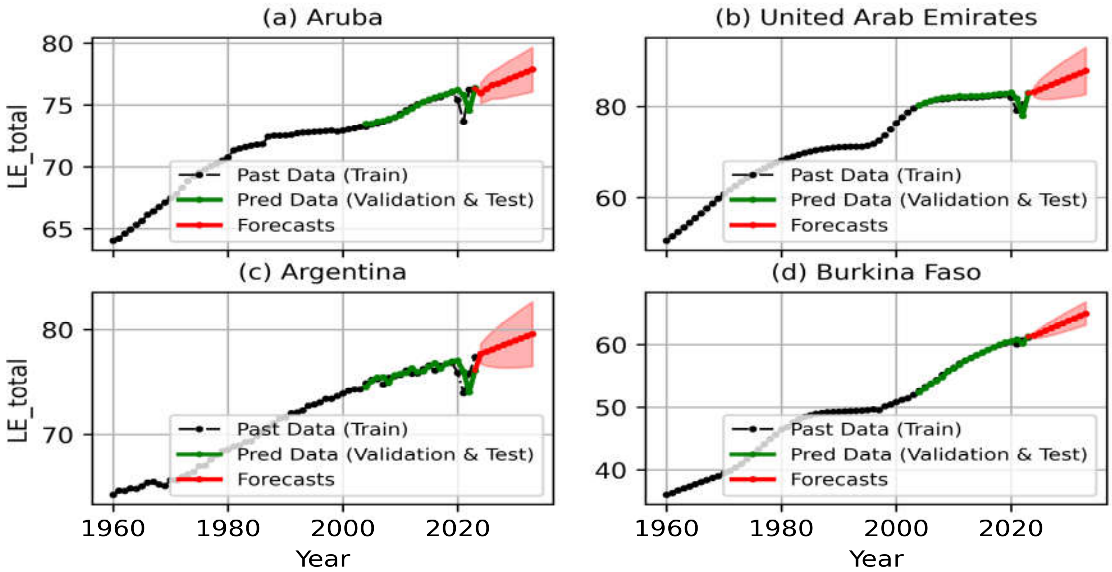


Figure 11

The forecasts of health expenditures using optimum model (countries: 5 to 8)

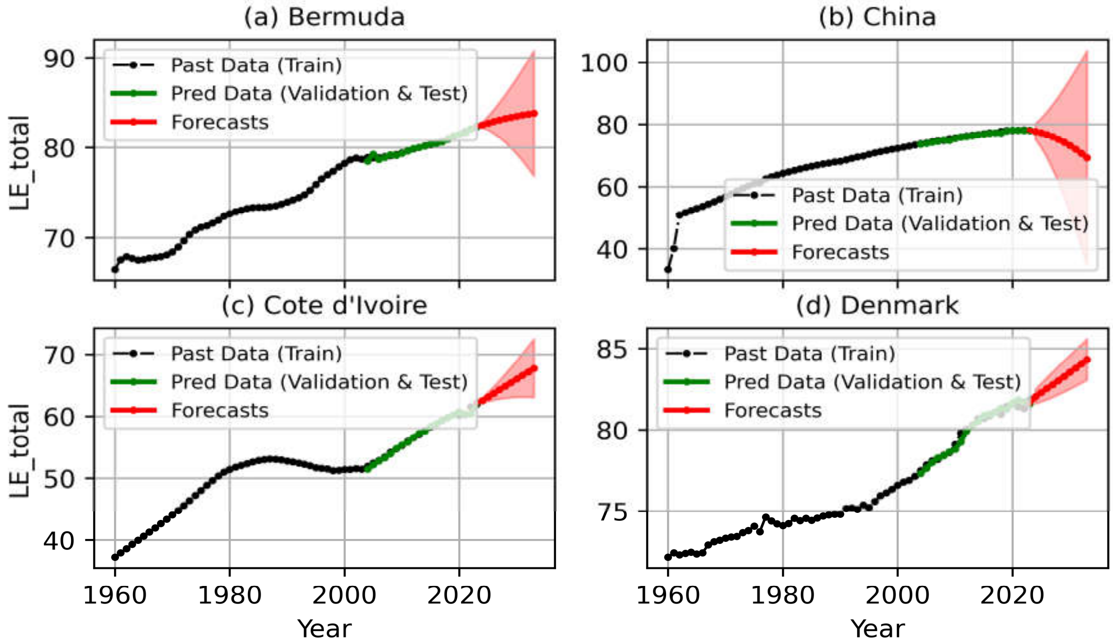


Figure 12

The forecasts of health expenditures using optimum model (countries: 9 to 12)

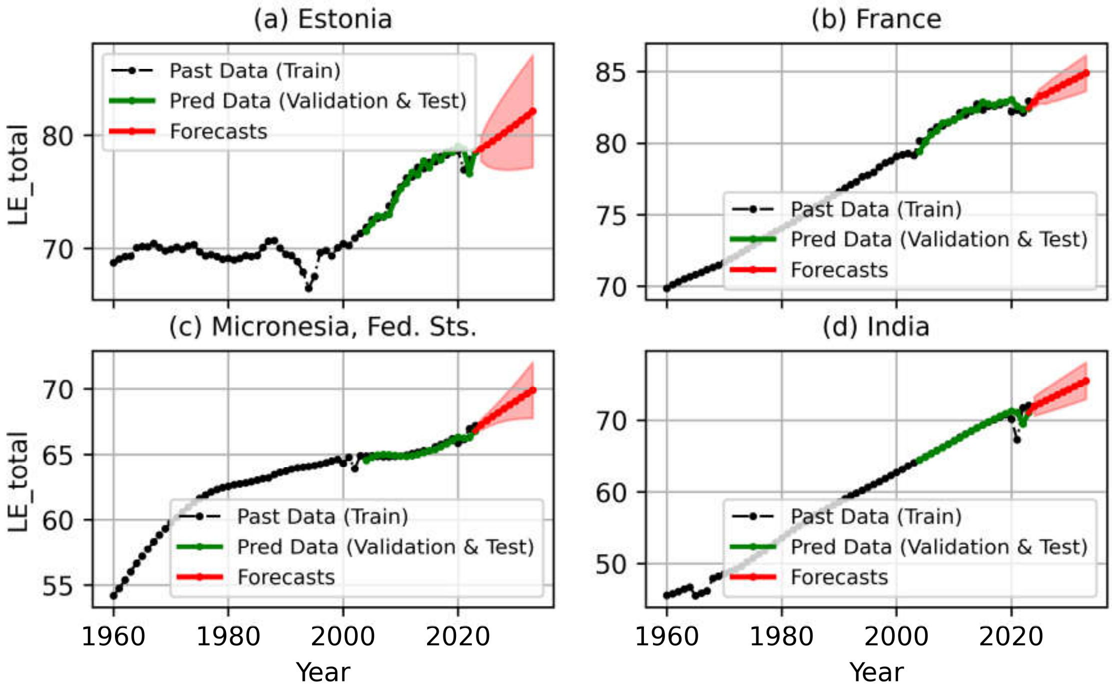


Figure 13

The forecasts of health expenditures using optimum model (countries: 13 to 16)

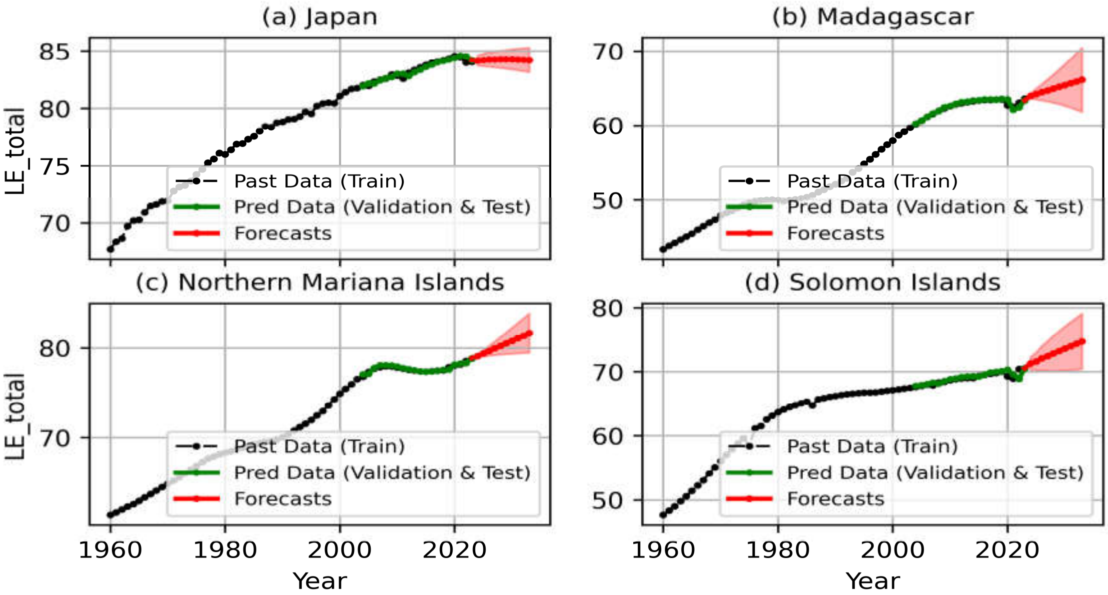


Figure 14
The forecasts of health expenditures using optimum model (countries: 17 and 18)

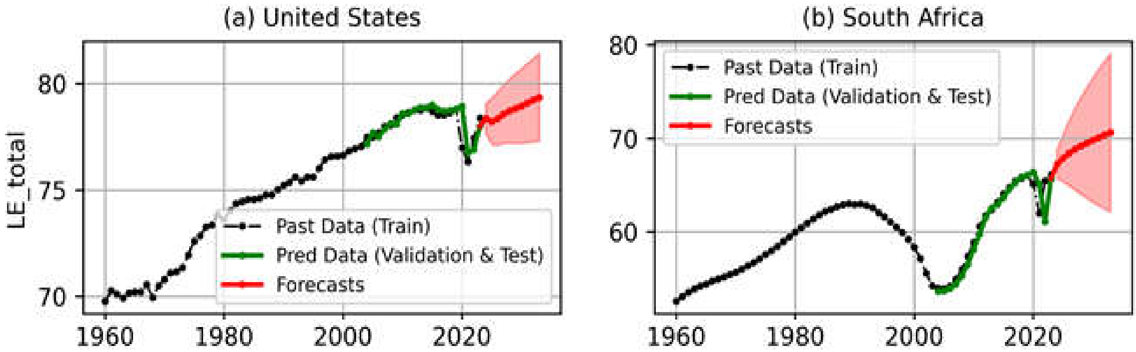


Table 6
Error metrics of selected SARIMA model

Country	MAE	RMSE	PMAE
Aruba	0.4134	0.7938	0.0055
United Arab Emirates	0.6758	1.0870	0.0084
Argentina	0.6657	0.9387	0.0088
Burkina Faso	0.1631	0.2789	0.0027
Bermuda	0.0441	0.0519	0.0005
China	0.1621	0.2051	0.0021
Cote dIvoire	0.2122	0.3751	0.0035
Denmark	0.2259	0.2807	0.0028
Estonia	0.5725	0.7304	0.0074
France	0.2955	0.3677	0.0036
Micronesia, Fed. Sts.	0.2177	0.2798	0.0033
India	0.6655	1.2856	0.0096
Japan	0.2039	0.2421	0.0024
Madagascar	0.2111	0.3211	0.0034
Northern Mariana	0.0773	0.1072	0.0010
Solomon Islands	0.3650	0.5643	0.0052
United States	0.3350	0.5980	0.0040
South Africa	0.8820	1.5530	0.0140

4. Conclusion

In the present study, the Fourier transform assisted ARIMA model is fitted to the historic global annual life expectancy data of the period 1960-2023. Fourier transform estimates into the periodicities of the timeseries. The model is evaluated using 5-fold cross validation using validation data for the years 2016 to 2023 are used for validation and testing. Using the best fitted ARIMA model with respect to AIC, the timeseries is predicted for the test period, and then error measures MAE, MAPE and RMSE are computed. The forecast of annual life expectancy for the period 2024 to 2029 is carried out. From the visualization plots the forecast appears to be in line with the patterns or trends present in the historical data. It is shown that the proposed method can be successfully used for predicting or forecasting data. It is conjectured that the proposed Fourier transform assisted ARIMA model performs better for reasonably longer data lengths, if the model order is appropriately selected.

5. Future Scope of Work

The limitations of the study are the assumption of homoscedasticity, i.e., uniform variance throughout the range of the timeseries variable, limited number of samples (64) and short-term predictions. By considering a generalized autoregressive conditional heteroscedastic (GARCH) model may provide better prediction and forecasting performances. Vector ARIMA models and Deep learning techniques like LSTM model can also be explored for better forecasting of life expectancy, especially for long-term predictions.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

[1] George Box E.P., et. al. (2016). *Time Series Analysis: Forecasting and Control*. 5th ed. John Wiley & Sons, Inc.

[2] Ruey S. Tsay. (2010). *Analysis of Financial Time Series*. 3rd ed. John Wiley & Sons, Inc.

[3] William W.S. Wei. (2019). *Multivariate Time Series Analysis and Applications.* 1st ed. John Wiley & Sons Ltd.

[4] Kitagawa, G., Takanami, T., and Matsumoto, N. (2004). State Space Approach to Signal Extraction Problems in Seismology. *In: Brillinger, D.R., Robinson, E.A., Schoenberg, F.P. (eds) Time Series Analysis and Applications to Geophysical Systems*. The IMA Volumes in Mathematics and its Applications, vol 139. Springer, New York

[5] Perrott MH, Cohen RJ. (1996). An efficient approach to ARMA modeling of biological systems with multiple inputs and delays. *IEEE Trans Biomed Eng*., 43(1), 1-14.

[6] Bahman Moghimi , Abolfazl Safikhani , Camille Kamga , and Wei Hao. (2018). Cycle-Length Prediction in Actuated Traffic-Signal Control Using ARIMA Model. *Journal of Computing in Civil Engineering*, 32(2), 1-10

[7] T.E. Arijaje et al. (2025). Modelling and Prediction of Satellite Signal Path Loss using the ARIMA models at Ku-band in Lagos State, South Western Nigeria. IOP Conference Series: Earth and Environmental Science. 1-13

[8] Schutgens, N. A. J. (2008). Simulating Range Oversampled Doppler Radar Profiles of Inhomogeneous Targets. J. Atmos. Oceanic Technol., 25, 1514–1528.

[9] Rabiner L, Biing-Hwang, and Yegnanarayana. B (2008). *Fundamentals of Speech Recognition*, Pearson Education, New Delhi.

[10] Venkata. KRao M. (2013) Rao MVK, "Machine analysis and synthesis of spoken Telugu vowels," *Third International Conference on Computational Intelligence and Information Technology*, Mumbai, 93-104.

[11] Maddela, V.K.R., Bhaskararao, P. (2022) Phonetic–Acoustic Characteristics of Telugu Lateral Approximants. Circuits Syst Signal Process, vol 41, pp.3508–3546, Feb 2022

[12] Krao MV and B Rao Peri. (2021). Durational and Formantshift characteristics of Telugu alveolar and bilabial nasal phonemes", MysuruCon-2021*, IEEE Mysore Subsection Flagship International Conference*, 386-393,

[13] Maddela, V.K.R., Bhaskararao, P. (2022) Phonetic–Acoustic Characteristics of Telugu Lateral Approximants. *Circuits Syst Signal Process*, 41, 3508–3546

[14] Fattah, Jamal et. al. (2018). Forecasting of demand using ARIMA model. International Journal of Engineering Business Management.

[15] Mukhairez, Hosam & El-Halees, Alaa. (2018). Medium-term Forecasting for City Water Demand and Revenue. *International Journal of Intelligent Computing Research*. 9 (2), 913-920.

[16] Abotaleb, M and Makarovskikh. T. (2021). System for forecasting covid-19 cases using time-series and neural networks models", *Engineering Proceedings*, 5(1), 1-8.

[17] Schaffer, A.L., et. al. (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. *BMC Med Res Methodol* **21**, 58

[18] Swaraj A, et. al. (2021). Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India. *J Biomed Inform.* 121:103887, 1-11.

[19] Li, WenQiang & Zhang, Chang. (2022). Application of Combination Forecasting Model in Aircraft Failure Rate Forecasting. Computational Intelligence and Neuroscience.

[20] Mubang, Ng, K.W., et. al. (2023). Experimental evaluation of baselines for forecasting social media timeseries. *EPJ Data Sci.* 12, 8

[21] Nkongolo, M. (2023). Using ARIMA to Predict the Growth in the Subscriber Data Usage. *Eng*, *4*(1), 92-120.

[22] Ospina, R., et. al. (2023). An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil. *Mathematics*, *11*(14), 3069.

[23] Mahmud Hasan. et. al. (2023). ARMA model development and analysis for global temperature uncertainty. *Frontiers in Astronomy and Space Sciences*, *10*, 1-11.

[24] Chiang, S., Zito, J., et. al. (2024). Time-series analysis. In Statistical Methods in *Epilepsy*, 166-200.

[25] Dhawani Shah and Manishkumar Thaker. (2024). A Review of Time Series Forecasting Methods. *International Journal of Research and Analytical Reviews,* 11(2).

[26] John Wang,, et.al. (2024). A fusion of machine learning algorithms and traditional statistical forecasting models for analyzing American healthcare expenditure. *Healthcare Analytics*, *5*, 1-10.

[27] Ewurum N.C. and Okafor S. O. (2024). Global Health Security: An Evaluation of Public Health Expenditure, Health Status and Economic Development in Nigeria. Sustainable Development., In *Global Health Security - Contemporary Considerations and Developments,* Eds. Allincia Michaud, et. al. IntechOpen, Jun. 19, 2024.

[28] Pereira da Veiga, C., *et al.* (2024). Implementation of the ARIMA model for prediction of economic variables: evidence from the health sector in Brazil. *Humanit Soc Sci Commun,* 11:1038, 1-10.

[29] Agaj, T., Budka, A., Janicka, E. *et al.* (2024). Using ARIMA and ETS models for forecasting water level changes for sustainable environmental management. *Sci Rep* **14**, 22444.

[30] Al Sulaie, S. (2024). Use of ARIMA Model for Forecasting Consequences Due to Traffic Crashes in the Kingdom of Saudi Arabia. *Journal of Road Safety*, *35*(4), 54–65.

[31] Sinulingga, Wita & Purba, Ronsen & Pasha, Muhammad Fermi. (2025). Combination of Regression and ARIMA Methods (Reg–ARIMA) Stock Price Prediction Model. *Journal of Comp. Net., Arch. and High-Performance Comp,* 7, 329-340.

[32] Zou, Yiyang. (2025). Forecasting Apple Inc. Stock prices: A comparative analysis of ARIMA, LSTM, and ARIMA-LSTM models. Advances in Operation Research and Production Management. 4. None-None. 10.54254/3029-0880/2025.23870

[33] Eissa, Mostafa. (2025). Modeling microbiological counts in purified water at a healthcare facility using ARIMA. Quantum Journal of Med. and Heal. Sci., 4. 56-68.

[34] Oikonomou, K., Damigos, D. (2025). Short term forecasting of base metals prices using a LightGBM and a LightGBM - ARIMA ensemble. *Miner Econ* 38, 37–49

[35] VK Rao, Maddela. (2025). Comprehensive Review of Nonstationary Seasonal Timeseries Models through Python Simulations. *Journal of Systems Engineering and Electronics*, 35(7), 64-70.

[36] David Muñoz-Rodríguez, et. al. (2025). Innovation ARIMA models application to predict pressure variations in water supply networks with open-loop control. Case study in Noja (Cantabria, Spain), *Energy Nexus*, 18.

[37] VK Rao, Maddela, (2025). Global Health Expenditure Forecasting using Fourier Transform assisted SARIMA models. IEEE- *2nd Int. Conf. on Int. Algo. for Comp. Intelligence Sys.* IEEE conference record #65746, (Presented at conference; To appear at IEEE Xplore)

[38] Rob J. Hyndman, Anne B. Koehler, J. Keith Ord and Ralph D. Snyder. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Series in Statistics. Springer-Verlag Berlin Heidelberg.

[39] Rob J. Hyndman et. al. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *Int. J. of Forecasting*, 18, 439-454.

[40] Robert H. S., and David S.S., *Time Series Analysis and Its Applications with R examples*, 4th ed., Springer Texts in Statistics, 2016.

[41] Rob J. Hyndman and Khandakar Y. (2008). Automatic Time Series Forecasting: The forecast Package for R, *Journal of Statistical Software*, 27(3), 1 – 22

[42] Athanasios Papoulis and S Pillai. (2019). *Probability Random Variables and Stochastic Processes*. 4th ed. McGraw Hill.

[43] ---"GHE: Life expectancy and healthy life expectancy", The Global Health Observatory, World Health Organization, 2023 https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy