Stochastic Modelling and Computational Sciences

A COMPARATIVE STUDY OF GMDH-TYPE ANN MODEL AND ARIMA MODELS FOR THE FUTURE PREDICTION OF ASFR AGE GROUP 25-29 IN INDIA

Minakshi Mishra[1], Anuj Kumar[2], Sumit Kumar[3*] and Bhagchand Meena[4]
[1, 2]Department of Statistics, School of Physical and Decision Science, Babasaheb Bhimrao Ambedkar University (Central University), Lucknow-226025, India
[3]Department of Mathematics, Chandigarh University, Mohali, Punjab, India
Corresponding author- stats.sumitbhal@gmail.com
[4]Department of Statistics, Central University of Rajasthan. NH-8
Kishangarh, Ajmer, Rajasthan, 305817

ABSTRACT
The present study aimed to compare the group method of data handling type artificial neural network (GMDH-type ANN) model and autoregressive integrated moving average (ARIMA) models, for the future prediction of India's age-specific fertility rate (ASFR) age group 25-29. In this study, time series data concerning the ASFR age group 25-29 was collected from the period 1995–96 to 2019–2020. The mean square error (MSE), root mean square error (RMSE) mean absolute error (MAE), mean absolute percentage error (MAPE), scatter index (SI), Akaike's information criterion (AIC) and Bayesian information criterion (BIC): have been used to compare the performance of various considered models. The GMDH-type ANN model is superior to another conventional statistical model i.e. ARIMA models. Thus, the GMDH-type ANN model was used for the prediction of the ASFR age group 25-29 over the next 20 years. The government will use the information regarding the prediction of ASFR age group 25-29 to allocate incoming resources and plan for children's care.

Keywords: ASFR, ARIMA, GMDH-type ANN

1. INTRODUCTION
India's fertility rate has significantly declined, reaching replacement levels in 18 of its 29 states [8]. This reduction has been driven by a variety of factors, with marriage, contraception, and abortion playing key roles [21]. The decline in fertility has had both positive and negative effects, including a reduction in maternal mortality and a shift in the population pyramid [16]. However, there is still a need for improvements in sexual and reproductive health services, particularly in terms of access and quality. Despite these challenges, there is potential for significant economic returns from family planning investments, particularly if the fertility rate can be further reduced [6].

[3] uses statistical analysis to examine the patterns in birth rates in India, offering valuable information about the efficacy of government initiatives related to family planning. [17] examines the demographic transition in India and the difficulties arising from its increasing population size and ongoing increased fertility and death rates. These variables might have had an impact on the birth rate in the year 2020. [9] found that an ARIMA (5,1,1) model had the highest accuracy in forecasting the decrease in birth rates in Tamil Nadu.

According to [4] the primary measure of fertility is the GFR. It is determined by dividing the total number of live births in a specific geographic area (such as a nation, state, or county) by the female population aged 15-49 (usually estimated for a mid-year). The resulting fraction is then multiplied by 1000. The GFR is often used as a measure of overall fertility because it represents the population with the greatest risk of giving birth. It is often based on freely available data that includes both the numerator and denominator, covering a large age range that includes most of the female reproductive years. This makes it a widely used and reliable measure of fertility [15]. The GFR in India is 67, with an urban GFR of 53.7 and a rural GFR of 73.7, as per the latest data report from 2020.

Several researchers considered using ARIMA models for the future prediction of fertility rates. [18, 22, 23, 24]. ARIMA models are not appropriate for long-term future prediction of non-linear data such as ASFR age group

25-29 when using the GMDH-type ANN model. The GMDH-type ANN model has been applied in various fields, demonstrating its versatility and effectiveness. introduced a dynamic GMDH-type ANN model for system identification, while [19] proposed a hybrid GMDH and Box-Jenkins model for time series forecasting, with the GMDH model outperforming other methods. [10] further enhanced the GMDH-type ANN model by incorporating evolutionary algorithms, improving performance. These studies collectively underscore the potential and practicality of the GMDH-type ANN model in various domains. A comparative study was conducted by researchers to evaluate the GMDH-type ANN technique in comparison to ARIMA and Holt's techniques [1, 13]. The researchers determined that the GMDH-type ANN technique has greater prediction capabilities compared to other models.

This study aims that make future predictions of ASFR age group 25-29 for India using the GMDH-type ANN model and this model has recently garnered significant interest and finds extensive applicability in many real-world problems.

2. MATERIALS AND METHODS

The data for the period 1995 to 2020 were collected from the source ‘‘indiastat.com”, MS-Excel, R, and GMDH-type ANN shell Software were used for the data analysis.

2.1 Autoregressive Integrated Moving Average (ARIMA)

The ARIMA models, created by [5] are a classic econometric technique for time series forecasting. The autoregressive moving average (ARMA) model is modified by the ARIMA. An ARMA model expressed the conditional average of Y_t as a function of both previous observations $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$, and previous innovations, $\varepsilon_{t-1}, \varepsilon_{t-q}$. The number of previous observations that Y_t depends on, p, is the AR degree. The number of previous innovations that Y_t depends q, is the MA degree. In general, these methodologies are denoted by ARMA (p, q). The form of the ARMA (p, q) approach is

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \ldots + \varphi_q \varepsilon_{t-q} \ldots \quad (1)$$

where α is a constant term, $\beta_1, \ldots, \beta_p$ autoregressive (AR) coefficients, φ_1 moving average (MA) coefficients, $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$, AR lags corresponding to non-zero, $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ MA lags corresponding to nonzero, MA coefficients, and degree of differencing D, if D has value 0 which means no integration. $\beta(B)Y_t = \varphi(B)\varepsilon_t$ where B is the backward shift operator, an estimate of the ARIMA model**.**

The ARIMA methodology, developed by Box and Jenkins in 1970, is used to determine a class of linear time series models. The Box-Jenkins approach's linearity restriction has been solved by statisticians in many different kinds of methods, and in addition to nonlinear time series models, robust versions of several ARIMA models have been developed [12]. In this work, forecasting and fitting were done using a box-Jenkins' ARIMA model technique.

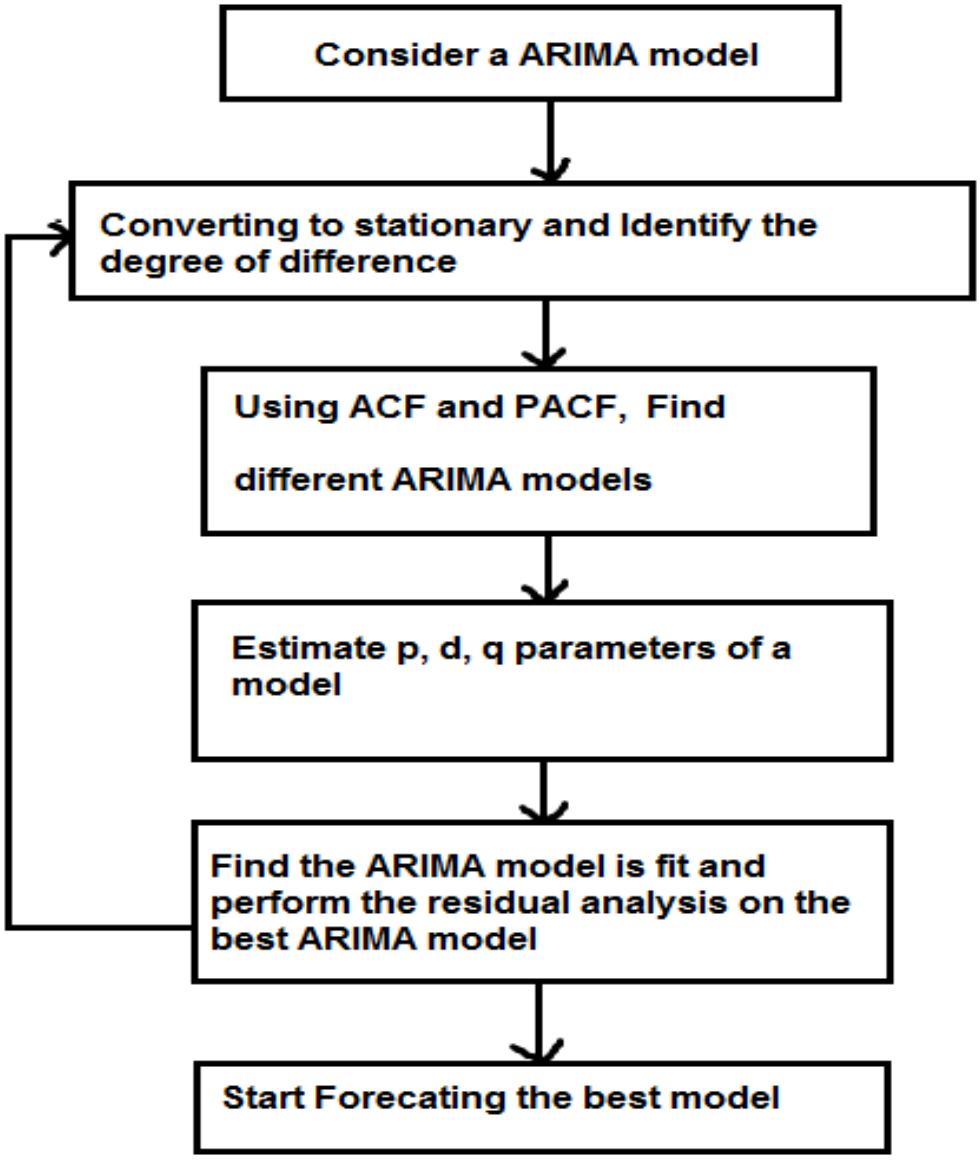

Figure 1. Structure of the Box-Jenkins methodology

2.2 Group Method of Data Handling Type Artificial Neural Network (GMDH-type ANN)

The GMDH method was first used by [7] to look at complex structures made up of a set of data with multiple inputs and a single output. For the GMDH network, the main goal is to create a feed-forward network function using a second-degree transfer function. The GMDH method naturally picks the best model structure based on the input factors, hidden layer neurons, and layer count.

$$\hat{y} = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k + \ldots\ldots \quad (2)$$

where x represents the input variable, a_i represents coefficients, y represents the output variable and m represents the number of observations of input variable. Figure 2 shows the process of developing the GMDH-type ANN model.

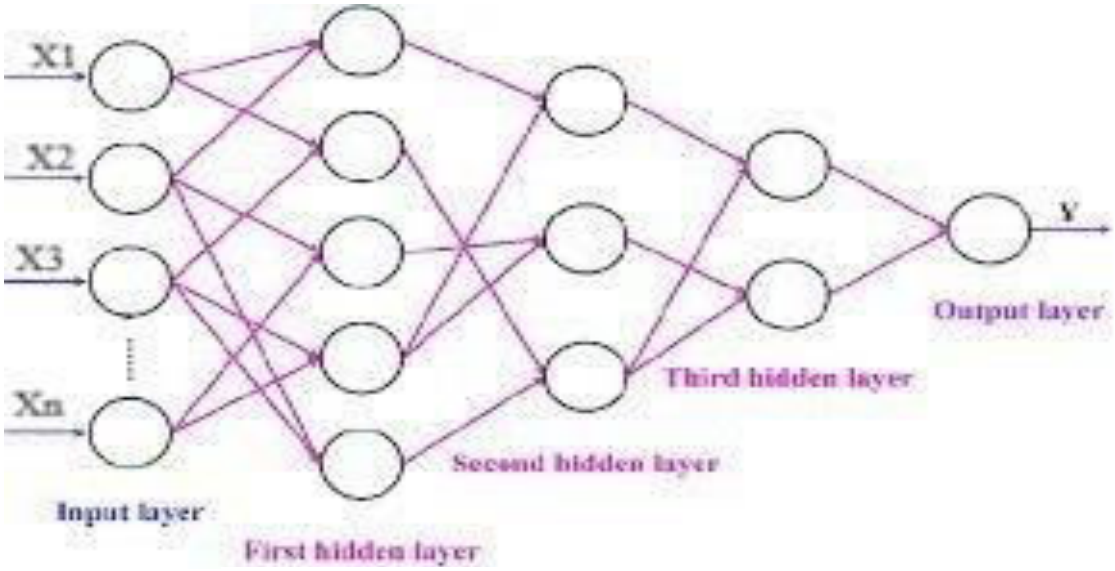

Figure 2. GMDH- type ANN model development process

2.3 Prediction Accuracy Measures

Model selection criteria are the standards by which a model's performance for the data being studied is evaluated to ascertain if a consistent pattern in the models' performances is present [2, 11, 20]. The accuracy measures used to determine and validate the results of various ARIMA models for predicting ASFR age group 25-29 in India include, root mean square error (RMSE), mean absolute error (MAE), Akaike's information criterion (AIC) and Bayesian information criterion (BIC). In the case of the GMDH-type ANN model, accuracy measures such as mean square error (MSE), RMSE, MAE, mean absolute percentage error (MAPE), and Scatter index (SI). The SI is a normalized measure of error, frequently presented as a percent. Lower values of the SI are an indication of superior model performance [14].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

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

$$SI = \frac{RMSE}{\bar{X}}$$

AIC = $2 \times K - 2 \times \ln L$

BIC = $K \times \ln N - 2 \times \ln L$

Where

N= is the number of observations

$\hat{y}_i$ = Predicted value of y

y_i = The actual value of y

$\bar{X}$ = Mean of actual value of y

K = The number for the parameters in the model

L = the likelihood function

3. RESULTS AND DISCUSSION

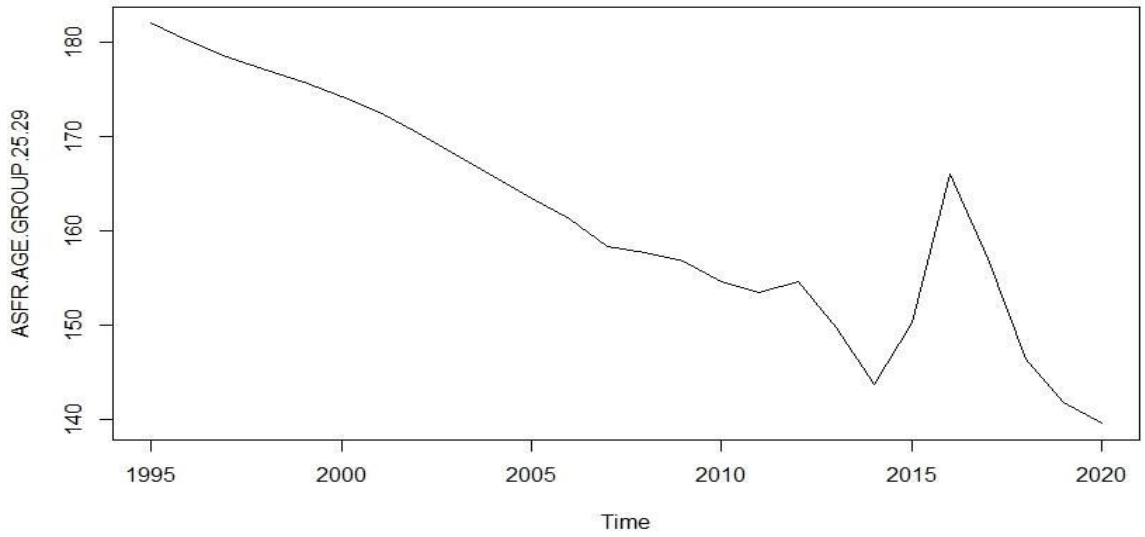

Figure 3. Graphical presentation of ASFR age group 25-29 of India (1995-2020)

Figure 3. shows the declining trend and the non-stationarity of the given period for ASFR age group 25-29 of India (1995-2020).

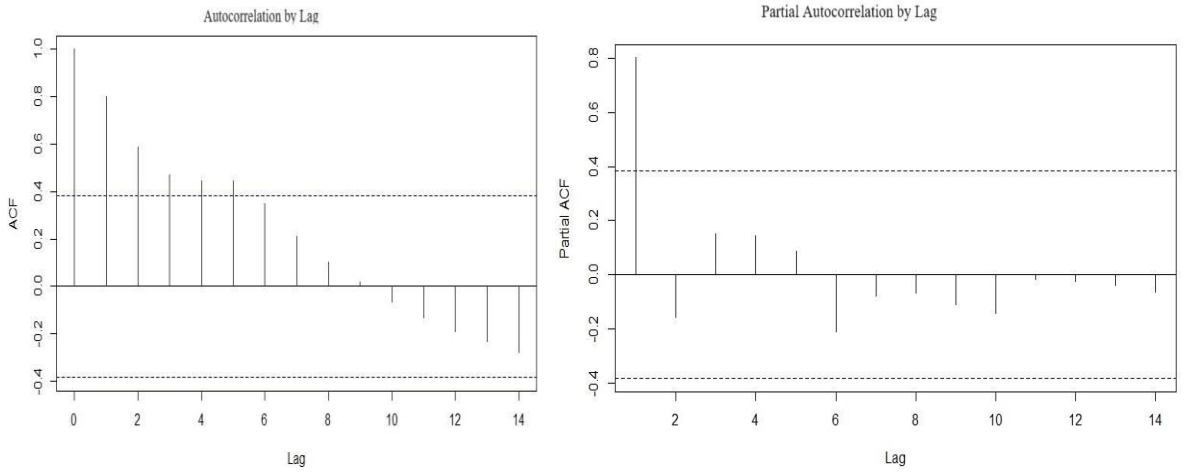

Figure 4. ACF and PACF at the Lag= 14 for ASFR age group 25-29 of India (1995-2020)

Plots for ACF and PACF are shown in Figure 4 accordingly. The autocorrelation values show a progressive decline from the initial second-order autocorrelation coefficient to the final value. The table shows that the Augmented Dickey-Fuller (ADF) test confirmed the acceptance of the Null hypothesis, indicating that the ASFR age group 25-29 of India (1995-2020) follows non-stationarity.

Differencing the data series of ASFR age group 25-29 to reach the stationarity:

The Box-Jenkins method requires time series data that is stationary. A correlogram graphic is used to assess the stationarity of time series data. The stationarity of the data's time series was evaluated. If stationarity is not attained, the differencing method is used to eliminate variations in the series, resulting in a consistent mean and variance throughout time.

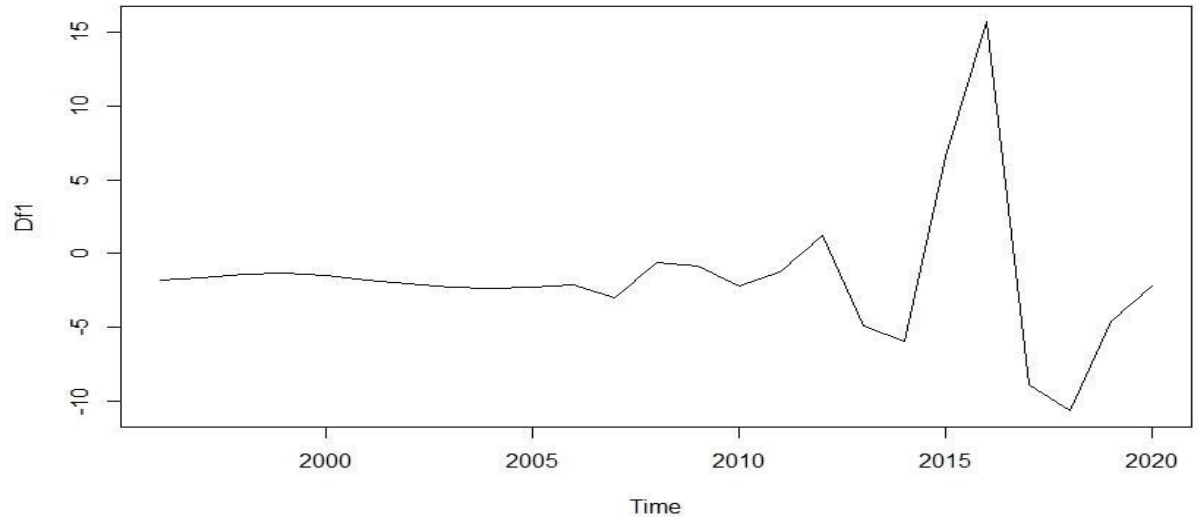

Figure 5. Plot the first difference of the time series data for ASFR age group 25-29.

The figure 5 shows that the ASFR data is given first-order differencing to create non-stationarity. A transferred series after second-order differencing is shown in Figure 6. The second-order differencing to the ASFR age group 25-29 data to make stationarity, and the ACF and PACF plots after second-order differencing are shown in Figure 7 respectively.

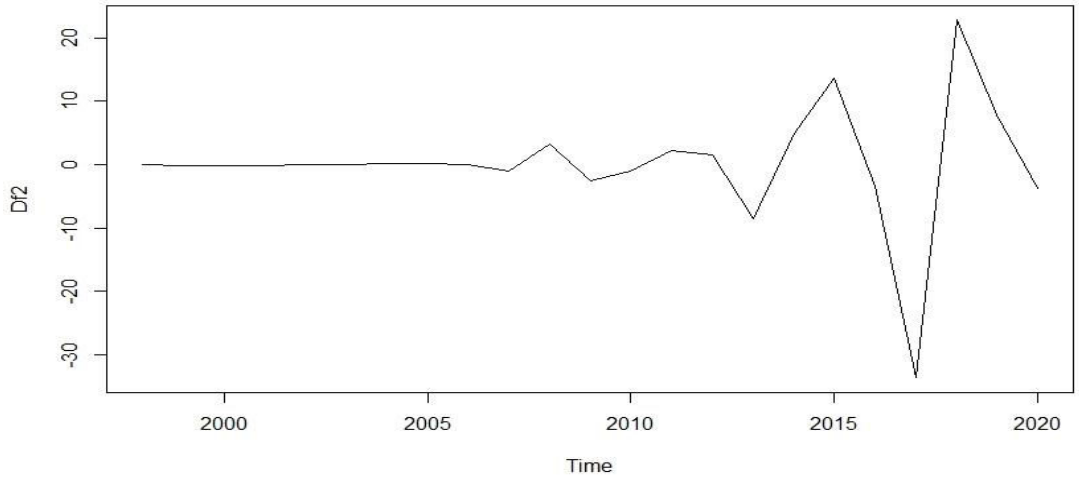

Figure 6. Plot the second difference of the time series data for ASFR age group 25-29.

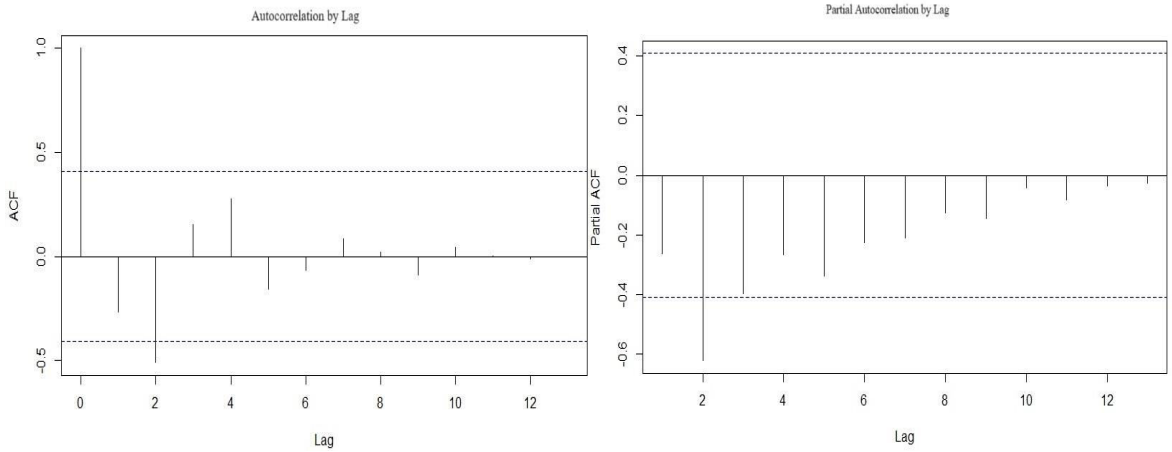

Figure 7. ACF and PACF plots for the second differencing of the ASFR age group 25-29 series.

The ADF test provides a test statistic and a p-value as outcomes. The test statistic is compared against critical values at significance levels, typically set at 5%. If the p-value is below the significance level, you can reject the null hypothesis and determine that the time series is stationary. When the p-value is above the significance level, you accept the null hypothesis and conclude that the time series has a unit root.

Table 1. ADF to check the stationarity of ASFR age group 25-29 of India

Parameters	**Time series data**	**1st order differencing**	**2nd order differencing**
ADF -Test statistic	-2.2232	-3.0025	-7.0380
Lag order	2	2	2
p-value	0.4874(NS)	0.1905(NS)	**0.01(S)**
Level of Significance	0.05	0.05	0.05

Note: Significant (S): p-value <0.005, Non- significant (NS): p-value >0.005

Table 1 shows that stationarity is generated by the 2nd-order differentiated data in ASFR age group 25-29 at a significance level of 5%. The alternative hypothesis is accepted when the p-value is less than 0.05.

The ARIMA models' accuracy is determined by four measures: RMSE, MAE, AIC and BIC values are used to choose the most suitable model for the transformed data.

Table 2. ARIMA Model selection criterion for ASFR age group 25-29

Models	RMSE	MAE	AIC	BIC
ARIMA (1,1,1)	4.3670	2.7847	152.01	155.66
ARIMA (1,1,2)	4.2462	3.0012	152.67	157.54
ARIMA (1,1,3)	4.0563	2.8358	152.97	159.06
ARIMA (1,1,4)	3.4628	2.2777	151.39	158.70
ARIMA (1,1,5)	3.4984	2.3423	152.85	161.38
ARIMA (1,1,6)	3.4093	2.2682	154.83	164.57
ARIMA (1,2,1)	4.5772	2.3996	151.90	155.42
ARIMA (1,2,2)	**4.2045**	**2.1822**	**149.90**	**154.61**

The ARIMA (1,2,2) model is the most suitable for the second-order differenced data of the ASFR age group 25-29 in India, as seen in Table 2. The top-performing model in this study, ARIMA (1,2,2), was selected by evaluating measures like RMSE, MAE, AIC and BIC where it showed the lowest values compared to all other models. The ARIMA (1,2,2) model follows a normal distribution with a mean of 0 and a variance of 1.

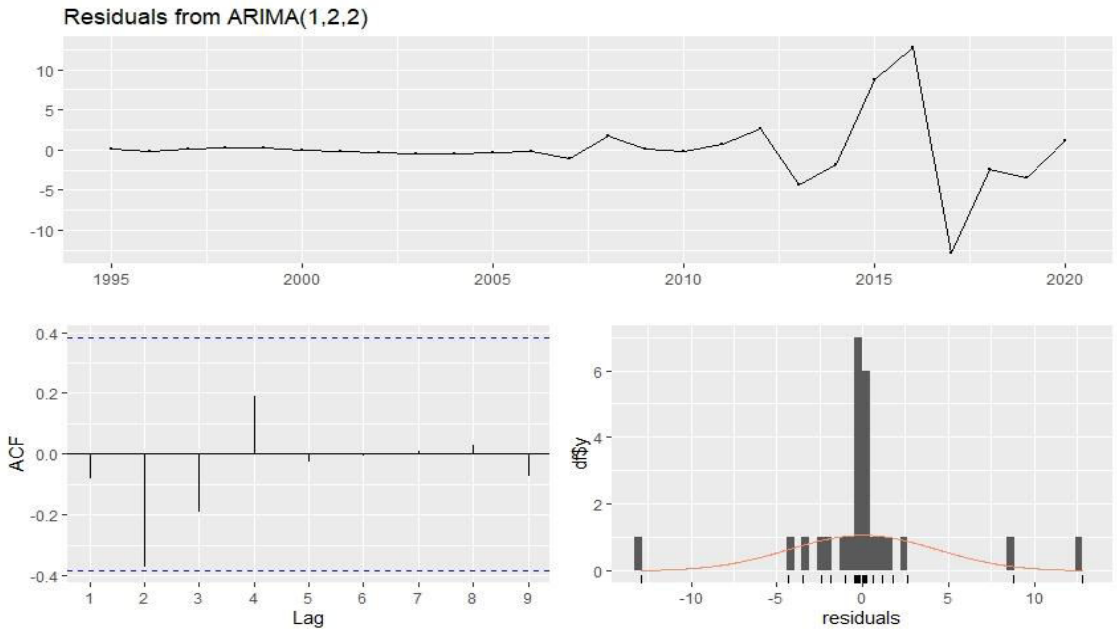

Figure 8. Residuals plot in ARIMA (1,2,2) model for ASFR age group 25-29

Figure 8, from the ARIMA (1,2,2) model for India's ASFR age group 25-29 that were statistically significant were included in the results of the study. A normality test was used to determine the normality of the residuals.

Model Comparison

The prediction's assessment of accuracy is a key issue when choosing Criteria. However, measures like RMSE, MAE, MAPE, and SI are used to measure prediction accuracy, displayed in Table 3. For the Indian ASFR age group of 25 to 29, the model with the lowest values of these measures provides an appropriate and consistent.

Table 3 Comparative performance metrics between the ARIMA (1,2,2) model and the GMDH-type ANN model for ASFR age group 25-29 in India

Models	ARIMA	GMDH-type ANN
Best parameters	**p = 1, d = 2, q = 2**	**Training set=80%, Testing set= 20%**
RMSE	4.2045	**0.2129**
MAE	2.1822	**0.1488**
MAPE	1.4041	**0.0937**
SI	0.0260	**0.0013**

As shown in Table 3 the SI value (0.0013) was the lowest for the GMDH-type ANN model as compared to the ARIMA (1,2,2) model. Further comparison between the GMDH-type ANN model and the ARIMA (1,2,2) model with RMSE, MAE, and MAPE indicated that the GMDH-type ANN model's performance was better as all error measures depicted lower values (0.2129, 0.1488, and 0.0937).

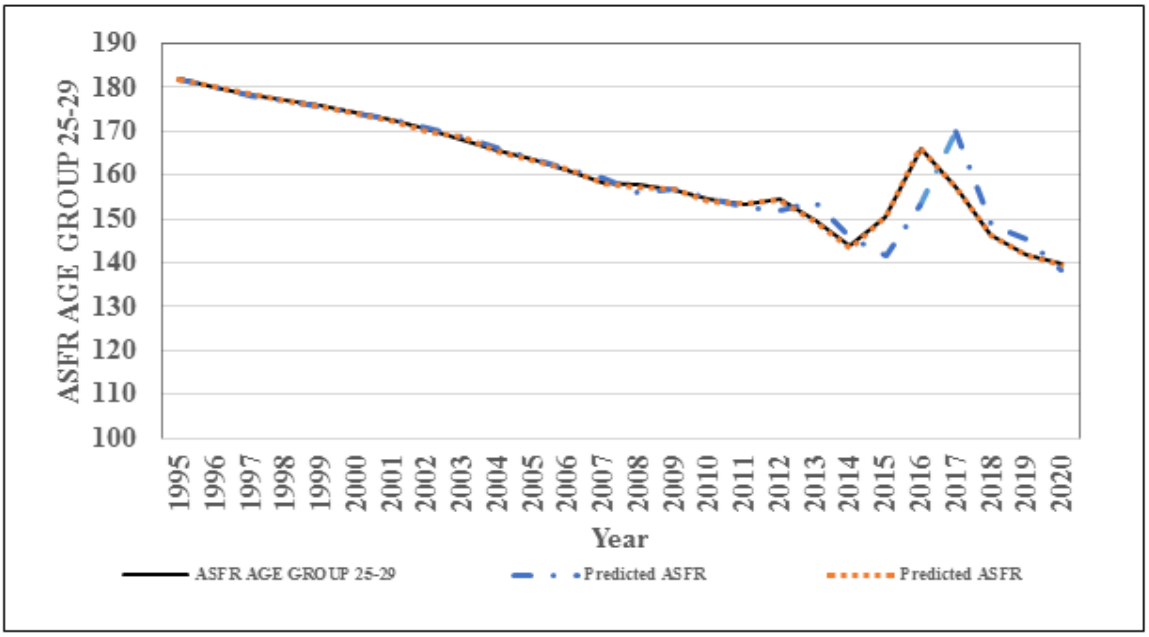

Figure 9. shows the Actual ASFR age group 25-29 and predicted ASFR age group 25-29 in India using two modeling techniques and data from 1995 to 2020.

The ASFR age group of 25-29 has shown a regular decline from 1995 to 2020. The ARIMA and GMDH-NN models accurately predicted the post-rates for the ASFR age group 25-29 from 1995 to 2020, similar to the actual rates (figure 9).

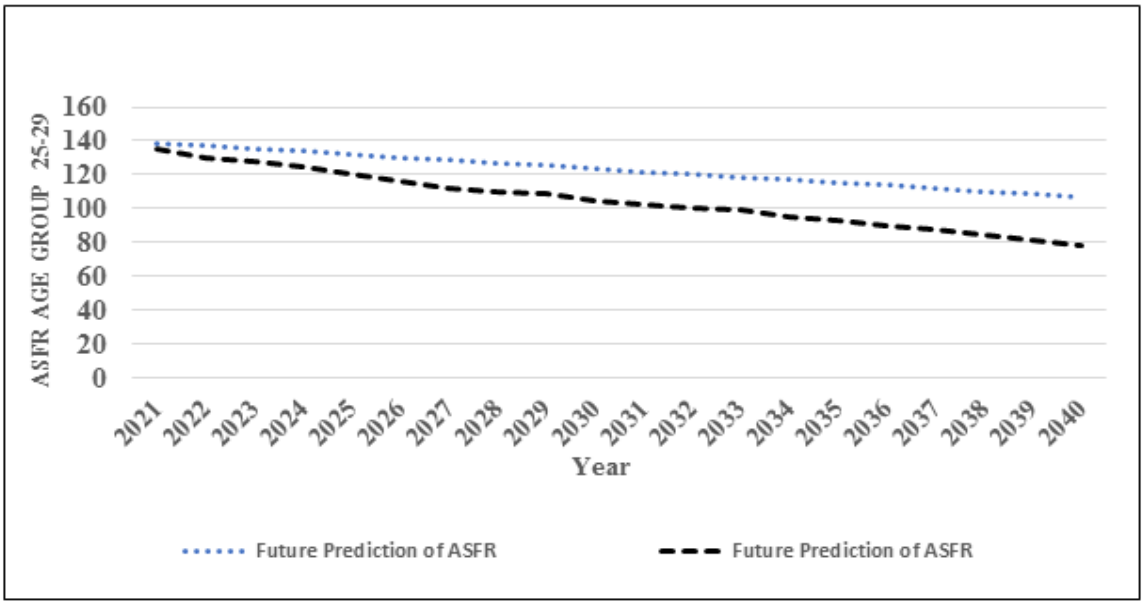

Figure 10. The future prediction for 2021 to 2040 is the ARIMA and GMDH-type ANN models.

Figure 10 clearly shows that the ARIMA and GMDH-type ANN models have significant differences in their ability to predict future outcomes from 2021 to 2040, based on sample data. The GMDH-type ANN model's future prediction is better than others, coming in at 78 (95% prediction interval 88.9524 - 68.9495) per 1000 live births by 2040.

4. CONCLUSION

The study shows that the GMDH-type ANN model performs better than the ARIMA (1,2,2) model in ASFR age group of 25-29 for accurate prediction in India, compared to other conventional ARIMA models. The GMDH-type ANN model is particularly suitable for analysing non-linear or unpredictable distribution data, such as fertility rates. The government will use the future fertility rate to allocate incoming resources and plan for children's care. To improve the fertility rate in the nation, the Indian government has started the use of several kinds of contraceptive techniques.

ACKNOWLEDGMENT

The authors are highly thankful to the Editor-in-Chief for his immense timely support and helped revise the manuscript with efficient reviewers' comments.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

1. Adeyinka, D. A., & Muhajarine, N., (2020). Time series prediction of under-five mortality rates for Nigeria: comparative analysis of artificial neural networks, Holt-Winters exponential smoothing and autoregressive integrated moving average models. *BMC Medical Research Methodology,* 20(1), 1-11.
2. Bal, C., Demir, S., & Aladag, C. H. (2016). A comparison of different model selection criteria for forecasting EURO/USD exchange rates by feed forward neural network. *International Journal of Computing, Communication and Instrumentalism Engineering*, *3*, 271-275.
3. Bhadha, M. (2020). Statistical analysis of birth rate in India. *International Journal of Advance Research, Ideas and Innovations in Technology, 6*, 159-161.

4. Bhende, A. A., & Kanitkar, T., (1997). *Principles of Population Studies*. Himalaya Publishing House, New Delhi, India.

5. Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco, Holden-Day.

6. Goli, S., James, K. S., Singh, D., Srinivasan, V., Mishra, R., Rana, M. J., & Reddy, U. S. (2023). Economic returns of family planning and fertility decline in India, 1991–2061. *Journal of Demographic Economics*, *89*(1), 29-61.

7. Ivakhnenko, A. G. (1968). The group method of data handling, a rival of the method of stochastic approximation. *Soviet Automatic Control*, *13*(3), 43-55.

8. Jejeebhoy, S., Santhya, K. G. and Zavier, A. F. (2020). Sexual and reproductive health in India. In *Oxford Research Encyclopedia of Global Public Health*.

9. Karunanidhi, D., & Sasikala, S. (2023). Robustness of predictive performance of arima models using birth rate of Tamilnadu. *Journal of Statistics Applications and Probability*, *12*(3), 1189-201.

10. Kim, D., & Park, G. T. (2005, June). GMDH-type neural network modeling in evolutionary optimization. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 563-570). Berlin, Heidelberg: Springer Berlin Heidelberg.

11. Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2022). On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, *73*(5), 937-954.

12. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting methods and applications*. John wiley & sons.

13. Mishra, M., Kumar, A., & Thakur, R. K. (2023). Superiority of gmdh neural network model over holt's and arima models for the future prediction of general fertility rate: a case study of india. *Advances and Applications in Mathematical Sciences*, *22* (3), 769-778.

14. Mentaschi, L., Besio, G., Cassola, F., & Mazzino, A. (2013). Problems in RMSE-based wave model validations. *Ocean Modelling*, *72*, 53-58.

15. Pathak, K. B., & Ram, F., (2016). *Techniques of Demographic Analysis*. Himalaya Publishing House, New Delhi, India.

16. Radkar, A., (2020). Indian Fertility Transition. *Journal of Health Management*, 22(3) 413–423.

17. Ram, U., & Ram, F. (2021). Demographic Transition in India: Insights Into Population Growth, Composition, and Its Major Drivers.

18. Salis, M. T. P., & Bramantoro, A. (2022, August). Forecasting the demand of birth control pills using arima-garch. In *Proceeding International Conference on Information Technology, Multimedia, Architecture, Design, and E-Business*, *2*, 208-217.

19. Shabri, A., & Samsudin, R. (2014). A hybrid GMDH and box-jenkins models in time series forecasting. *Applied mathematical sciences*, *8*(62), 3051-3062.

20. Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. E. (2013). A survey of forecast error measures. *World applied sciences journal*, *24*(24), 171-176.

21. Singh, S., Shekhar, C., Bankole, A., Acharya, R., Audam, S., & Akinade, T. (2022). Key drivers of fertility levels and differentials in India, at the national, state and population subgroup levels, 2015–2016: An application of Bongaarts' proximate determinants model. *Plos one*, *17*(2), e0263532.

22. Waseem, H. F., & Yasmeen, F. (2023). Forecasting total fertility rates for urban and rural areas in Pakistan with a coherent functional model. *JPMA. The Journal of the Pakistan Medical Association*, *73*(7), 1440-1446.

23. Yifan W. (2023). Analysis of China's Birth Rate Prediction Based on Time Series. *Advances in Economics, Management and Political Sciences*, *55*, 205-217.

24. Yongjie, C. (2023). Model Analysis on the Birth Rate in China. *Advances in Economics, Management and Political Sciences*, *58*, 220-227.