Forecasting Female Literacy Rates in India Using Time Series Modelling

Article *in* Journal of Huazhong University of Science and Technology · April 2021

CITATION
1

READS
561

2 authors, including:

Aniket Biswal
VIT University
1 PUBLICATION **1** CITATION

SEE PROFILE

Forecasting Female Literacy Rates in India Using Time Series Modelling

[1]Aniket Biswal, [2]G. K. Revathi

[1]School of Electrical Engineering, Vellore Institute of Technology Chennai, Chennai-127

[2] Division of Mathematics, School of Advanced Sciences, Vellore Institute of Technology Chennai, Chennai-127

Abstract

India is a country with a tag of one of the world's fastest growing major economy from 2014 to 2018.Literacy is essential for economic development and social well-being. Our economy improves when students have high levels of learning. Therefore, it is very critical to help the planning commissions and different policy makers of the states to design and implement appropriate schemes to be able to bridge the gap in the literacy rates which have been present for several years. This study uses the gender-based literacy rates from 1881 – 2011 by Census of India 2011 conducted once in every 10 years. The study is used to predict the literacy rates of females using a very well popular time series modelling ARIMA modelling technique. The results indicate that the ARIMA(1,3,2) is the most suitable model to predict the female literacy rates with a 93.41% efficiency. Moreover, it can also be concluded from the results that female literacy of India will be 100% by 2031. This implies that the Indian government should continue making innovative schemes and policies to aid in decreasing the difference between the male and female literacy rates which have been decreasing since 1981.

Keywords

Female literacy, ARIMA model, Correlation, Box-Jenkins Approach, Regression Analysis, Augmented Dickey Fuller, Akaike's Information Criterion (AIC), AIC-corrected (AICc), Bayesian Information Criterion (BIC), forecasting.

1. Introduction

Literacy is the ability to read and write to comprehend information around us and communicate effectively according to our interpretation of the information. Even though literacy is at the foundation level of education, it is a crucial empowering tool for human development and thus the overall growth of a society and its economy.

The three key features of UNESCO's definition of literacy are-

1. Literacy is about the uses people make of it as a means of communication and expression, through a variety of media.
2. Literacy is plural, being practiced in particular contexts for particular purposes and using specific languages.
3. Literacy involves a continuum of learning measured at proficient levels.

Women add up to almost half the world's population. The best way to boost the health, nutrition and economic status of a family that make up for a diminutive part of a nation's economy. Hence, it can be stated that the lack of women's education can be barrier to a country's economic

development. In India women attain very less education compared to men. As per the report of the Census 2011, the gap in literacy between men and women is 11.2%.

Literacy constitutes the backbone of development in a progressing country like India.It enhances the quality of life, awareness, and skills of people. It is often considered as the first step towards liberty from socio-economic constraints, particularly for women, considering the constraints and stereotypes imposed on them. Independence is an important state for any individual, giving one a sense of stability and control and women today certainly require it. In order for women to achieve this state of independence, literacy certainly plays a role. It improves their self-esteem and increasing economic productivity allowing them to take control over their being. The development of a nation is depicted by how the women have thrived and hence it is essential to uplift and empower them. Indian government, to serve this purpose, has taken up a handful of initiatives to support women literacy. The National Scheme of Incentive to Girls for Secondary Education (NSIGSE) is offering Rs.3000.00 on behalf of unmarried girls under the age of 16 as a fixed deposit, with the right to withdraw the sum along with interest when they reach 18 years of age after passing grade X. Rastriya Madhyamik Shiksha Abhiyan (RMSA) aims to improve the quality of education by providing high school studies with a reasonable standard of living, improving quality of secondary education, removal of gender, social and economic barriers and disabilities.
Despite these reforms, women in India face challenges to be educated and a few of them are:

- Inadequate educational environment for girls: Unavailability of basic amenities like hygienic restrooms and drinking water is not encouraging for girls. Fewer female teachers also make it not preferable for parents to enrol girls to schools due to regards to their safety.
- Early marriage: Child marriage involves underage women, many of whom are in poor socio-economic conditions. These women are not enrolled to schools and are made to learn house holding activities.
- Poverty: Single biggest cause of illiteracy in India and others are majorly repercussions of it. In a poor family, women are the main victims of malnourishment and denied the opportunities of any sorts. Men are given the priority as that will be considered as an investment while on the other hand parents will not benefit from the girl child's education.

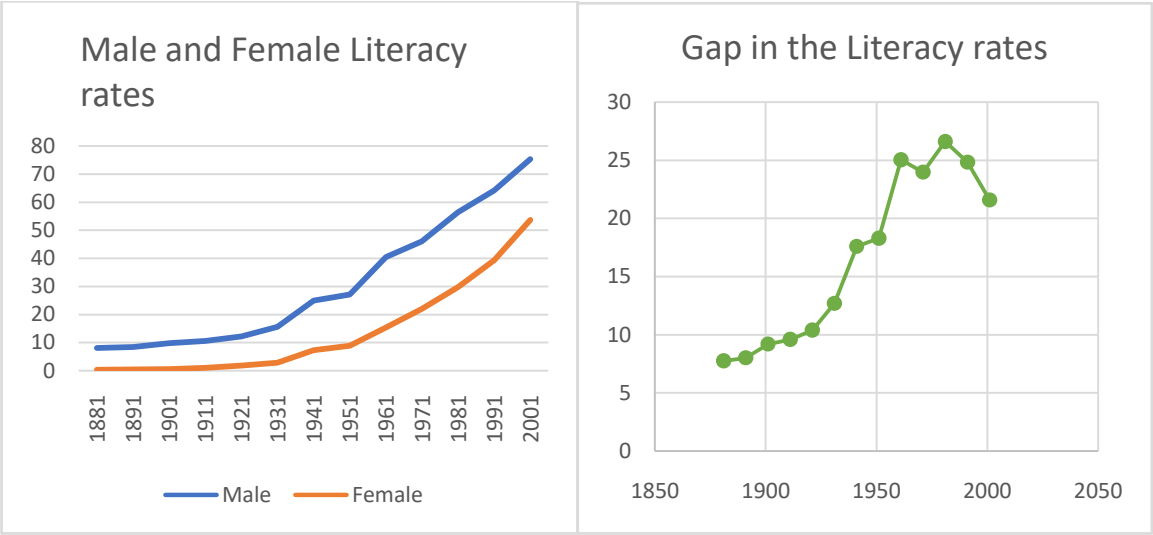

Figure 1: The plot of male and female literacy rates and the gap between them from 1881-2011

There are a lot of initiatives which are taken by the government both in term s of infrastructure and funding. At the same time a lot of improvements needs to be done. The gap in the literacy rates between male and female reached a peak of in 1981 with a difference in literacy of 26.62%. The literacy gap between male and female as you can see in the figure has been falling down since 1981 and has started to narrow down.

As highlighted earlier the significance of female literacy in a country like India, so it is very important to keep a track of literacy. It is the cornerstone of human development which in turn affects the growth of a country's economy, therefore it becomes important to improve the number of literates. Forecasting of literacy rate will help government & policy makers to develop strategies to improve literacy.

According to the Official journal of the Academy of Family Physicians of India, Female literacy is an important determinant of infant health and population stabilization. The effect of female literacy exists independently of male literacy. Therefore, it is an essential indicator required by governments as well as policy makers to make informed decisions which can be supported by the useful statistics obtained as a result of this extensive study. Moreover, from figure 2 below the literacy rates of male and female have a strong positive correlation with a Pearson's correlation coefficient equal to 0.98 and a p-value < 0.05 which indicates that female literacy rates can be used for our study to predict literacy rates in India.

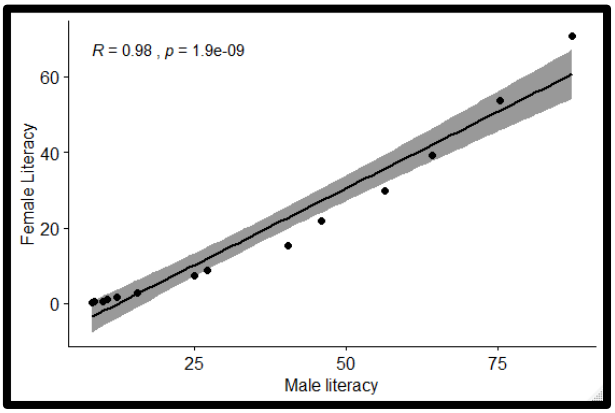

Figure 2: The scatter plot of male and female literacy rates along with Pearson's correlation coefficient R

The study uses time series ARIMA modelling to predict the female literacy rates in India. ARIMA modelling is a very well-known forecasting technique and is ideal when we are using a univariate time series which depends on its previous terms.

2. Method

In our study we will be using a very well-known time series forecasting model, **ARIMA** model which stands for **A**uto **R**egressive **I**ntegrated **M**oving **A**verage proposed by Box & Jenkins (1976), also known as the Box-Jenkins model. In theory, it is the most general class of models for forecasting a time series which can be made to be "stationary" by differencing (if necessary).A random variable that is a time series is stationary if its statistical properties are all constant over time. *A stationary series has no trend, its variations around its mean have a constant amplitude, and it wiggles in a consistent fashion*, i.e., its short-term random time patterns always look the same in a statistical sense.

A nonseasonal ARIMA model is classified as an **"ARIMA(p,d,q)"** model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of nonseasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.

In terms of y, the general forecasting equation is:

$$\widehat{y_t} = \mu + \varphi_1 y_{t-1} + \cdots \varphi_p y_{t-p} - \theta_1 e_{t-1} - \cdots \theta_q e_{t-q}$$

The forecasting equation is constructed as follows-

Let y denote the d[th] difference of Y, which means:

If d=0: $y_t = Y_t$

If d=1: $y_t = Y_t - Y_{t-1}$

If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

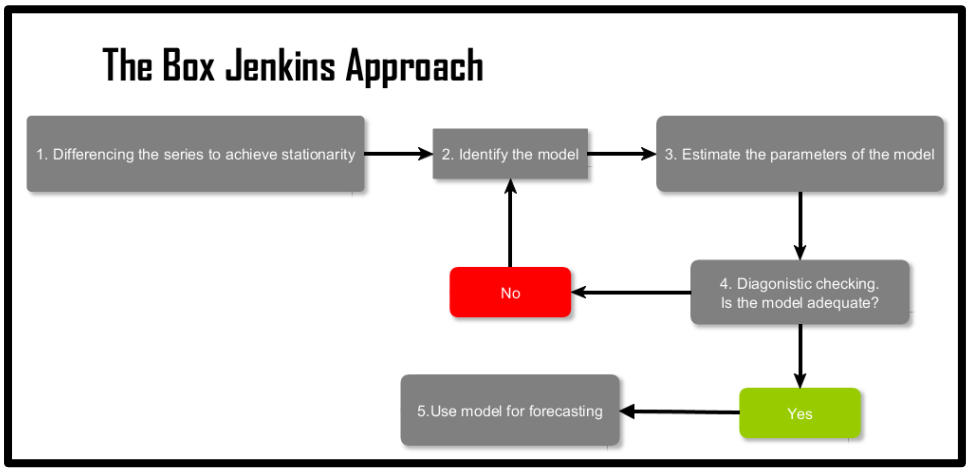

Figure 3: The Box-Jenkins Approach

The entire time series analysis and modelling can be done using the Box-Jenkins Approach as explained as follows–

Step 1: Data preparation

- Check whether the original time series data is stationary or not. This can be done using the augmented Dickey Fuller (ADF test). The ADF is common statistical test which belongs to the category of the 'Unit Root Test'. As the name suggests it is an augmented version of the Dickey Fueller test. It tests the null hypothesis assuming that a unit root is present and an alternate hypothesis assuming trend stationarity.
- Determine the appropriate values of d (number of differencing needed to make it stationary) and use ADF test to ensure the stationarity of the time series data.

Step 2: Identification/Estimation of ARIMA model

- This study used the values of Akaike's Information Criterion (AIC), AIC-corrected (AICc) and the Bayesian Information Criterion (BIC) to determine the best ARIMA model which has the lowest values of the three criteria. However, it is important to note that the temporary ARIMA models that could be used are based on the values of the autoregressive order (p), the moving average order (p) and the differencing process (d).

Step 3: Diagnostic Checking

- Check the residuals of fitted models.

Step 4: Forecasting

- Forecasting the future values using the model with the highest accuracy using indicators like mean error (ME), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE), and mean absolute percentage error (MAPE).

3. Results and Discussion

The dataset we have used for this study has female literacy rates having a frequency of a decade starting from 1881-2001 and we will use the 2011's data to test our model efficiency. The figure 4 below shows the rise in female literacy rates with a minimum of 0.35 in 1881 and maximum of 53.67 in 2001. However, from the figure you can clearly see that the rise between 1941 to 1951 i.e. from 16.1 to 18.33 is not as much as increase in rates as in other decades. The descriptive statistics for the literacy rates of females in India are given in the Table 1 below.

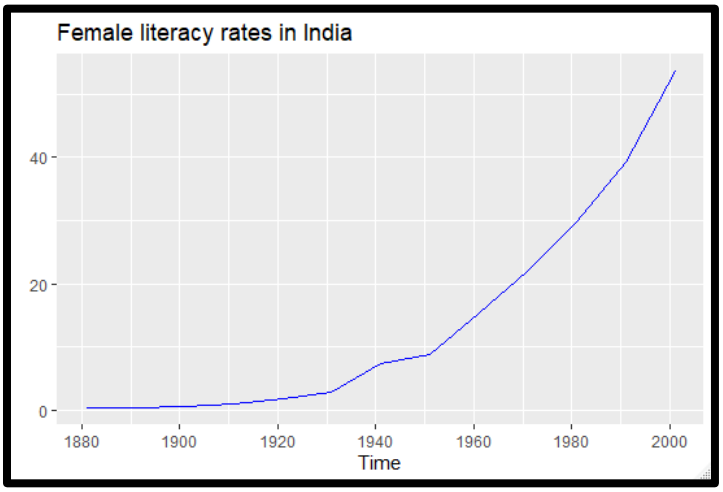

Figure 4: Female Literacy rates in India from 1881-2001

From figure 4, We can observe from the increasing trend of the data points. For a time-series data to be stationary as mentioned earlier, it should have statistical properties such as mean, variance, autocorrelation, etc. constant over-time. In figure 5, the plot of ACF for the original time series data is slowly decaying with time. If the ACF is slowly decaying, that means future values of the series are correlated / heavily affected by the past values. If past values of the series are high, the future values should also be high. Hence as the present value goes up, we can essentially say that the future values will also go up (assuming positive auto-correlation here). This means the mean will change over time, implying that the time series data is non-stationary. Therefore, differencing is required to make the data stationary. This is also verified using ADF test which is elaborated below.

Descriptive Statistics	
Mean	14.09769
Standard Error	4.796866
Median	7.3
Standard Deviation	17.29535
Sample Variance	299.129
Kurtosis	0.840151
Skewness	1.307875
Range	53.32
Minimum	0.35
Maximum	53.67
Sum	183.27
Count	13

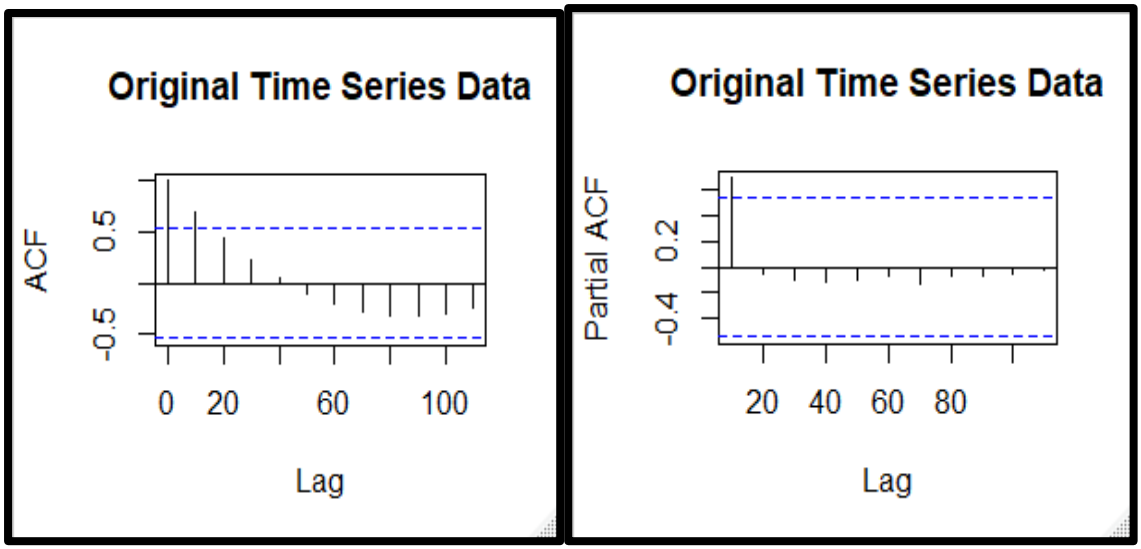

Figure 5: ACF and PACF plots of the original time series data (d=0)

On performing the Augmented Dickey-Fuller (ADF) test on the original time series data the value of Dickey-Fuller = 3.3391, Lag order = 2, p-value = 0.99. As the p-value is >5% (i.e. 0.05), differencing it once does not make the time series stationary. To make the data stationary we need to differentiate the data again and again perform the ADF test. On performing the ADF test after differencing, Dickey-Fuller = -6.3311, Lag order = 2, p-value = 0.01. The p-value is <5%, therefore the differencing of 2 i.e. **d=2** can be used in ARIMA (p,d,q) modelling. The model **ARIMA (p,2,q**) can be used for predicting the female literacy rates over the decades. The p and q order can be estimated or identified using the ACF and PACF plots obtained after appropriate order of differencing i.e. d required to make the series stationary. On differencing the series again i.e. d=3 and performing the ADF test, the Dickey-Fuller = -6.7492, Lag order = 2, p-value = 0.01. The lower the Dickey-Fuller value (more negative) the stronger the rejection of the null hypothesis (the series is non-stationary). Hence, the model **ARIMA(p,3,q)** with a Dickey-Fuller value of -6.7492 is more suitable than **ARIMA(p,2,q)** with a Dickey-Fuller value of -6.3311.And this can also be verified by their AIC as well BIC values of the models obtained while we move further in our study.

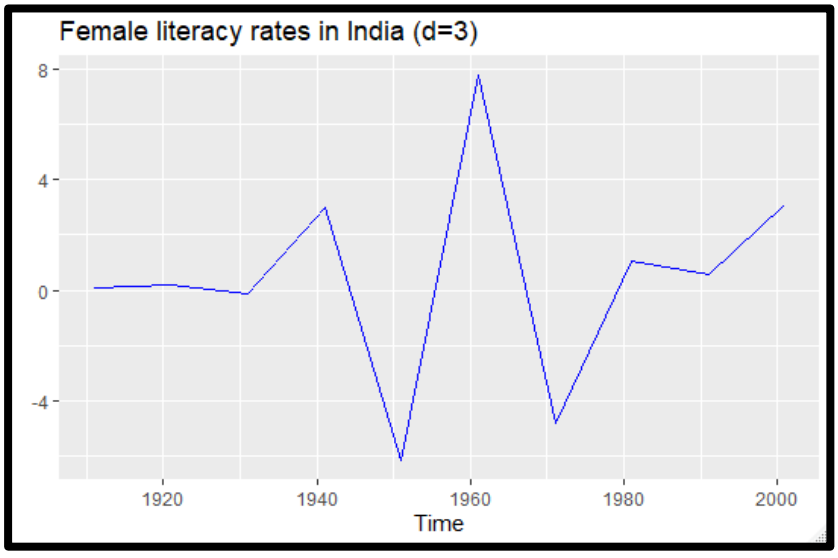

Figure 6, shows the ACF and PACF plots after differencing the time series data thrice i.e. d=3. The table 2 provides a guideline to choose model using the patterns in the ACF and PACF plots. In figure 6, the ACF plot dies out after the 2 lags cut-off which means MA(2) is suitable. At the same time the

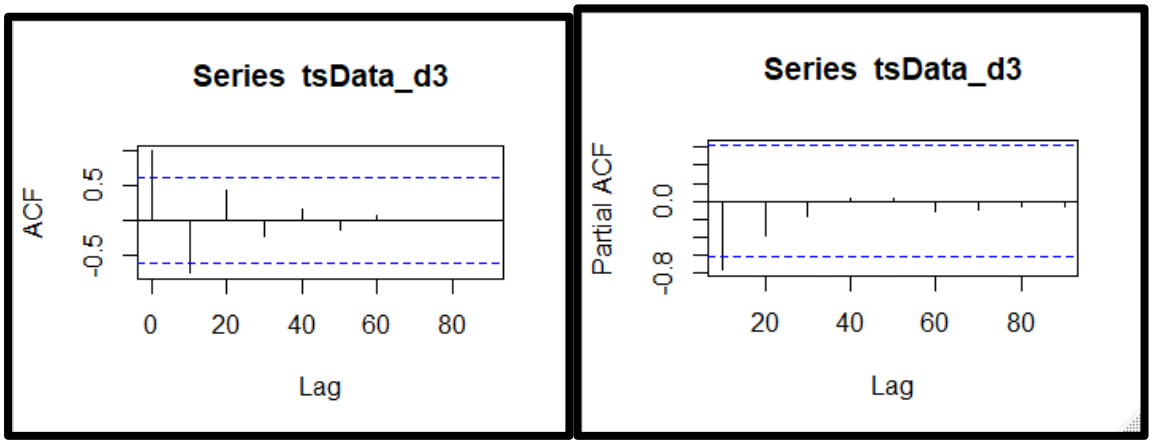

Figure 6: ACF and PACF plots after differencing d=3

PACF plot dies out after a single lag cuts-off which means AR(1) is suitable. Therefore, an ARIMA(1,3,2) model will be suitable for predicting the female literacy rates of India.

	Autocorrelations	Partial Autocorrelations
MA(q)	Cut off after the order q of the process	Die Out
AR(p)	Die Out	Cut off after the order p of the process
ARMA(p, q)	Die Out	Die Out

Table 2: Guidelines and patterns in the ACF and PACF plots to estimate p and q

There are a lot of alternate models which can be used to predict the female literacy rates by experimenting with different values of d as well as the order p and q which can be decided using the guideline table by looking at the patterns of the plots. The key lies here in validation of your model and whether your model gives the right and the most appropriate description of your data. The below table summarises the different alternate models which could be considered keeping in mind that the data was made stationary before applying ARIMA modelling to it.

Model	AIC	AICc	BIC	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA(1,3,2)	48.0994	52.0994	49.3097	0.4709	1.1842	0.7682	4.5160	6.5894	0.1729
ARIMA(2,3,2)	49.7245	57.7245	51.2374	0.4733	1.1078	0.6740	4.4966	5.7861	0.1517
ARIMA(2,3,3)	51.3868	66.3868	53.2023	0.4370	1.0859	0.6850	3.9940	5.9060	0.1542
ARIMA(0,2,2)	53.0812	54.5812	54.2749	0.8622	1.5894	1.0676	8.9555	11.2847	0.2403
ARIMA(0,2,0)	53.8100	54.2500	54.2100	1.1008	2.3453	1.5377	10.4714	15.4142	0.1091
ARIMA(2,2,3)	54.7915	58.2201	56.3831	0.8118	1.5860	1.0680	8.4465	11.3489	0.2404

Table 3: ARIMA models

In order to validate the best ARIMA model two criteria are used-

- *Akaike information criterion* (AIC)- estimates the quality of each model, relative to each of the other models.
- *Bayesian information criteria* (BIC)-is a variant of AIC with a stronger penalty for including additional variables to the model.

The lower the AIC/BIC value of the model the better fit it is. In this study the AIC criterion was used to find out the model which could be deployed to predict the literacy rates. The adequacy of the model should be checked by the randomness of the residuals of the model. The residual analysis is done on the model ARIMA(1,3,2) to check whether residuals are adequate for the fitted data.

Residual Analysis -

1. In figure 7, The**run sequence plot** shows that the residuals do not violate the assumption of constant location and scale. It also shows that most of the residuals are in the range (-1,2). Hence, fixed variation assumption is justified.
2. In figure 8, The **histogram** and **normal probability plot** indicate that the normal distribution provides an adequate fit for this model.

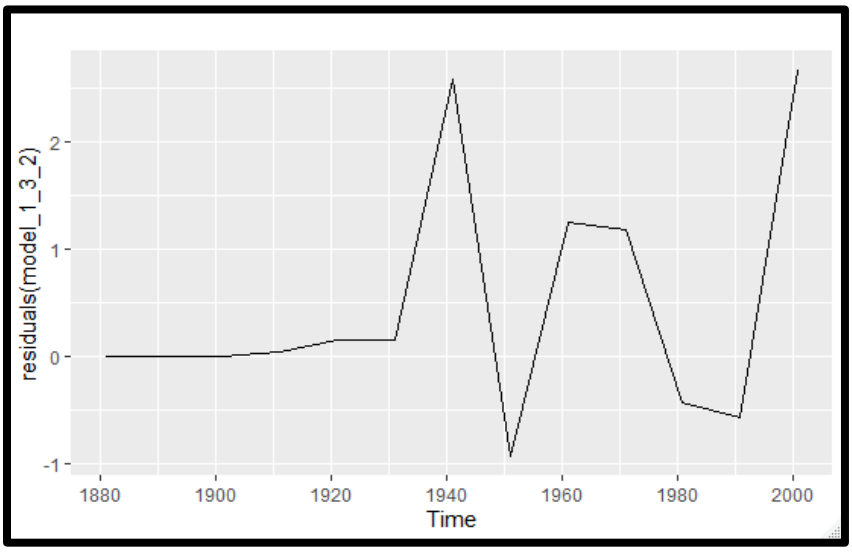

Figure 7: Run sequence plot of the residuals.

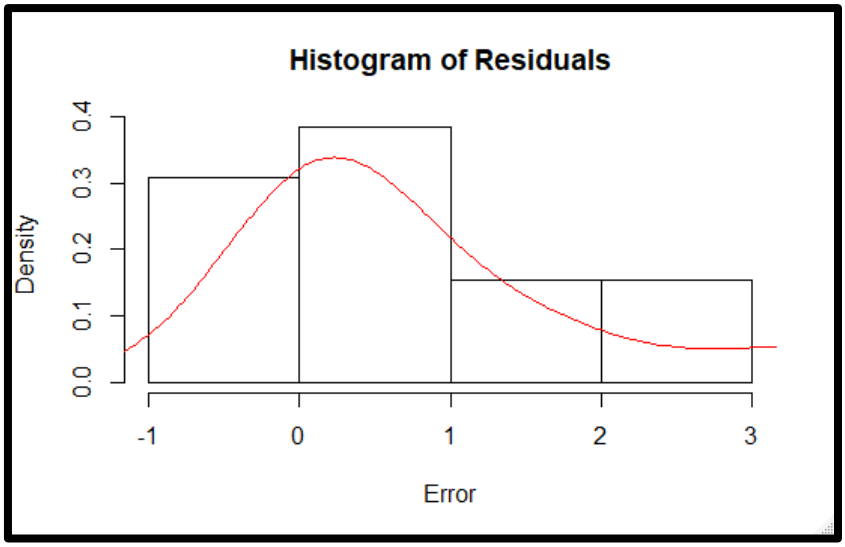

Figure 8: Histogram and normal distribution probability plot of the residuals.

The accuracy measures of the model to be used are equally important to decide the goodness of fit of the model. The following indicators can be used to find the accuracy each of them has different significance.

1. MAE (Mean Absolute Error) - It is the average of the absolute values of the error in the predicted data points.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

2. MASE (Mean Absolute Scaled Error) – It is a scale-free error metric that gives each error as a ratio compared to a baseline's average error.

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{1=2}^{n} |y_i - y_{i-1}|} \text{ and } \qquad MASE = |q_t|$$

3. RMSE (Root Mean Square Error) – It is one of the most widely used way to find accuracy. It compares between predicted value and actual value as described.

$$RSME = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)}{N}}$$

4. MAPE (Mean Absolute Percentage Error) -The mean absolute percentage error (MAPE) is a statistical measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

$$M = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{A_t - F_t}{A_t}\right|$$

Where A_t is the actual value and F_t is the forecast value.

Year	Point Forecast	Forecast Intervals			
		Lo 80	Hi 80	Lo 95	Hi 95
2011	67.1002	65.2106	68.9899	64.2103	69.9902
2021	84.6536	81.7064	87.6008	80.1463	89.1610
2031	104.0114	98.8203	109.2026	96.0723	111.9506

Table 4: Forecasted values i.e. point forecasts and forecasts for different confidence values.

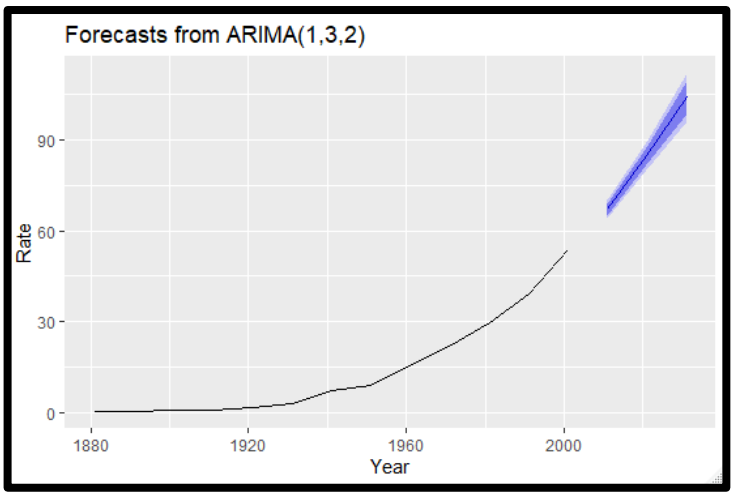

Figure 9.a: Forecasts of the female literacy rate in India using ARIMA(1,3,2)

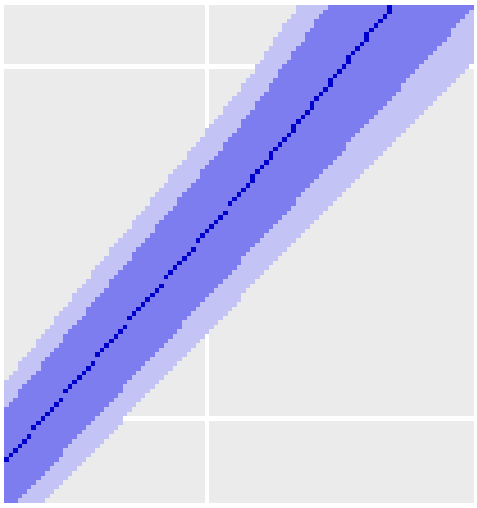

Figure 9.b: Confidence Limits

The figure 9 shows the plot of the forecasted values as obtained using the ARIMA(1,3,2) model. The blue line shows the actual point forecast predicted by the model. While the dark blue area is between Hi80 and Lo80 (i.e. 80% confidence limits) and the light blue area is between Hi95 and Lo95 (i.e. 95% confidence limits) as shown in the table. Over time the 95% and 80% confidence interval area increase as longer-term forecasts will be more vulnerable to have uncertainty, as the model will regress future Y on previously predicted values. The shape of the confidence bounds in this case, do not to widen with increasing horizon. This pattern in confidence bounds may signal that our model is a stable model.

In this article through ARIMA model it is found that on 2031 female literacy rate is expected to reach 100%. The correlation coefficient between male and female literacy rates is positive. Hence both of them depend on each other. When male literacy is increased gradually female literacy rate is also increased. The gap between male and female literacy rate is reducing.

4. Conclusion

The study in this article shows that 100% female literacy rate in India can be achieved by 2031 as predicted in this paper if Government continuously supports girls to pursue their education through some of the schemes mentioned like National Scheme of Incentive to Girls for Secondary Education (NSIGSE) and Rastriya Madhyamik Shiksha Abhiyan (RMSA). The results indicate that the ARIMA(1,3,2) is the most suitable model to predict the female literacy rates with a 93.41% efficiency. Through the suitable and novel government policies India will reach its 100% female literacy in the said year otherwise this can be achieved little late.

5. References

1. Pathak, S., & Gupta, A. (2013). Status of Women in India with Particular Reference to Gap in Male Female Literacy Rate in India. *International Journal of Environmental Engineering and Management*, *4*(6), 549-552.

2. Chandramouli, C., & General, R. (2011). Census of india 2011. *Provisional Population Totals. New Delhi: Government of India*, 409-413.

3. Chamdani, M., Mahmudah, U.R., & Fatimah, S. (2019). Prediction of Illiteracy Rates in Indonesia Using Time Series. *International Journal of Education, 12*, 34-41.

4. Mahmudah, U.R. (2017). Autoregressive Integrated Moving Average Model to Predict Graduate Unemployment in Indonesia. *Practice and Theory in Systems of Education, 12*, 43 - 50.

5. Ozaki, T. (1977). On the Order Determination of ARIMA Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 26*(3), 290-301. doi:10.2307/2346970

6. Olajide, J. T., Ayansola, O. A., Odusina, M. T., & Oyenuga, I. F. (2012). Forecasting the Inflation Rate in Nigeria: Box Jenkins Approach. *IOSR Journal of Mathematics (IOSR-JM)*, *ISSN: 2278*, *5728*(3), 5.

6. Jain, S., & Mishra, N.K. (2015). FORECASTING OF LITERACY RATE USING STATISTICAL AND DATA MINING METHODS.