

A Time Series Analysis of the Number of Female Examinees in Matriculation / HSLC Examination in Assam (India) Since 1951 and its Comparison with Male Counterpart

Geeta Rani Sarmah¹, Labananda Choudhury², Subrata Chakraborty³

¹Kamrup Academy Higher Secondary School, Assam, India

Email: geeta_12may@yahoo.co.in

²Department of Statistics, Gauhati University, Assam, India

Email: Lchoudhurygu@gmail.com

³Department of Statistics, Dibrugarh University, Assam, India

Email: subrata_arya@yahoo.co.in

*Corresponding Author: subrata_arya@yahoo.co.in

Received: 20th June 2016/ Revised: 20th November 2016 / Accepted: 11th March 2017

©IAppStat-SL2016

ABSTRACT

Completed secondary education of females plays a significant role in generating the opportunities and benefits of social and economic development. Therefore the knowledge of growth pattern of the number of female examinees in Matriculation / HSLC examination as an indicator of completion of secondary stage of education for effective budget and program planning is the need of the hour. Since independence, in Assam, a north-eastern state of India, the proportion of female examinees is increasing every year surpassing the number male examinees. The objective of the present study is to identify and fit a suitable time series model and forecast the number of female and male examinees both, using data for the period 1951-2015. Structural approach to analysis of time series data is adopted to construct several models, eliminating inappropriate ones and keeping the most suitable model. The selected models are validated in terms of its structure and forecasting accuracy. Our findings suggest that the ARIMA (1, 1, 0) and ARIMA (2, 1, 0) model as the best suited for forecasting of female and male examinees even when the outliers are detected and substituted through linear interpolation.

KEYWORDS: ARIMA; Autocovariance; RMSE; MAPE; MAE; Outliers.



1. Introduction

In India, since independence at the end of the secondary stage of schooling, i.e., at the end of tenth standard a public examination, commonly known as High School Leaving Certificate (HSLC) Examination or Matriculation Examination is conducted by different Universities/ Education Boards/Councils at state and national levels. Usually, this is the first public examination that a student encounters in his/her life and he/she is expected to be in the age group 15-16 years at the time of appearing in this examination. A reliable and valid examination can provide an equal opportunity for each examinee to show his/her level of competency and certify the completion of secondary education. Completed secondary education plays an important role in generating the opportunities and benefits of social and economic development (Secondary Education in India, 2009).

Assam is a state, located in the northeastern region of India. Here since pre-independence period, the appearance rates in Matriculation /HSLC Examination of male and female examinees varied widely. In 1951, 88% examinees were male and 12% were female. Their pass percentages were 41% and 51%, respectively. Again in 1956, among the examinees in this examination, 83% were males and 17% were females and so far as their pass percentages were concerned, it was 41% for male and 39% for female (Gauhati University Examination Results, 1948-1969). However with the passage of time this disparity in the appearance rate for males and females seems to be declining and along with it, a progressive trend is also seen in the pass percentages among both male and female examinees. This is to some extent reflected in the fact that in the year 1995, among the examinees in the HSLC examination 55% were males and 45% were females and their pass percentages were 38% and 28% respectively.

In the year 2010, 50% male and 50% female examinees did High School leaving Certificate Examination, their pass percentages being 67% and 60% respectively (Board of Secondary Education, Assam, 2010). While in the year 2015, there was 48% male and 52% female examinees and their pass percentages were 66.54% and 58.10% respectively.

The success achieved in education of females in India in the post-independence period is the result of a number of schemes forwarded by the government for bringing females to schools and retaining them there till they become eligible for appearing in the concluding examination of the secondary level of Education. Completed secondary education for females ensures that, they receive both the

benefit of primary education and additional benefits linked to further education (Secondary Education in India, 2009). The positive externalities of secondary education on health, gender equality and poverty reduction are stronger than those of primary education through its impact on young people's age at marriage, propensity to reduce fertility and improved birth practices and child bearing (Secondary Education in India, 2009). Expanded secondary education of female leads to significantly lower maternal and child mortality, slower population growth and improve education of children. One of the prime objectives of Rastriya Madhyamic Siksha Abhijan (RMSA) is the removal of sex related disparities in terms of appearance level in the Secondary level of Education. In 2002, World Bank estimated the social internal rate of return to secondary education to be 40% for females and 13% for boys. In 2004, private internal rate of return for females and boys were 26% and 15% respectively (Secondary Education in India, 2009). It is also observed by Rashtriya Madhyamik Siksha Abhijan (RMSA), Assam that enrollment of females has been higher than boys at the secondary level in schools across the state over the last six years. Data made available by the RMSA, Assam, reveals that between 2007 and 2012, the number of females enrolled in the secondary section, i.e., classes IX and X, went up constantly, surpassing corresponding number of boys. Though the number of boys enrolled at secondary level also increased in the last six years, it has been lower than that of females which reflects a positive trend in women's education in the state (The Times of India, December 3, 2013). At the World Summit in September 2005, governments convinced that "progress for women is progress for all" (ADB, UNDP, 2006). Under such situation provision for secondary education of relevant and good quality is a crucial tool for generating the opportunities and benefits of social and economic development.

At the World Education Forum, Dakar, 2000, countries agreed on ensuring that by 2015 all children, particularly females, will have access to complete free and compulsory education of good quality. In India focus on females' education was put in place by introducing several schemes such as in 1986 National Policy on Education, 1992 Program of Action, Sarva Siksha Abhijan (SSA) in 2001 followed by 2005 National Curriculum Frame-Work and Rashtriya Madhyamik Siksha Abhijan in 2009. One of the main objectives of RMSA is to ensure universal access to secondary education by 2017 and universal retention by 2020 (RMSA-India, 2012).

As we have entered in 2016, we should know the growth pattern of number of females appearing in HSLC examination as an indicator of completion of secondary stage of education for effective budget and program planning using time series data. A time series is a set of observations on the values that a

variable takes at different points of time. To reveal the growth pattern and to make best forecast of female examinees sitting in HSLC Examination in Assam using appropriate time series technique that can be able to describe the observed data successfully is need of the hour. To get a fare idea of gender disparity it is also useful to carry out a comparison of the male examinees with in the same reference period and parameters.

Among various time series forecasting techniques Autoregressive Integrated Moving Average (ARIMA) model pioneered by Box and Jenkins has been proved to be most powerful (Bisgaard and Kulachi, 2011). In the field of economics, finance, business, physical science, social science etc., Box-Jenkins techniques are extensively used to better understand the dynamics of a system and to make sensible forecasts about its future behavior (Bisgaard and Kulachi, 2011). However, the use of these techniques in the field of education for analyzing number of examinees in any examination is very limited. Education is a concurrent subject and it is state responsibility to contribute majority of expenditure at all levels of education, including secondary education (Aggarwal, 1993). Therefore accurate models may be important for the State as well as for the Nation for pursuing appropriate strategies for providing quality and relevant secondary and higher education to young generation for becoming active citizens and productive workers.

2. Objective of the study

The first objective of the present study is to propose a model of number of female examinees in Matriculation/ HSLC examination in Assam based on the data from 1951 to 2015 using ARIMA techniques to know the stochastic behavior of the study variable on their own under the philosophy “let the data speak for themselves” (Gujarati and Sangeetha, 2011). The second objective is to carry out the same exercise for male examines and contrast the findings. Final aim is to detect presence of outliers if any in the data and study the consequence of linear interpolation of these outliers in the model selection.

3. Sources of data

Time-Series data on female examinees in Matriculation/HSLC Examination from 1951 to 2015 in Assam were collected from different sources. Secondary education in Assam started in 1835. The class X public Examination was known by different names, like- Entrance Examination (conducted by Calcutta University till 1947), Matriculation Examination (conducted by Gauhati University from 1948 to 1963) and High School Leaving Certificate Examination (conducted by Board of Secondary Education, Assam since 1964) (The Assam Tribune, November 11, 2004). Therefore, data were collected from Gauhati University Information Center (1951 to 1963), District Library,

Guwahati (1964 to 1990) and Assam Board of Secondary Education (1991 to 2015). Closer examination of the collected data suggests an upward trend in the number of female as well as male examinees in Matriculation/HSLC Examination from 1951 onwards (see Figure 6), which is a reflection of positive trend in women education in the state. An irregular reduction in the number of female (male) examinees was observed during 1990-92.

It may be noted here that the examinees from recognized government or provincialized schools are termed as regular, who appeared from recognized private schools are termed as institutional private and appeared from non recognized private schools are called non institutional private examinees. It is observed that from 1985 to 2010 number of female (male) examinees in HSLC Examination from regular institutions was increasing over time while, the number of examinees from recognized and non recognized institutions was irregular in nature. This affected the nature of the time series under study.

It is pertinent to note here that the female (male) examinees in Assam High Madrassa, Central Board of Secondary Education (CBSE) and Indian Certificate of Secondary Education (ICSE) Examinations are excluded from the purview of the presented study.

4. Methodology

Assumption of time series forecasting is that the future depends upon the present while the present depends on the past (Chen, 2008). The main objective in time series modeling is to gain the ability to make forecast about the future on the basis of the selected model. The Box-Jenkins ARIMA (p, d, q) model is given as (Bisgaard and Kulachi, 2011)

$$z_t = \sum_{i=1}^{p+d} \varphi_i z_{t-i} + a_t - \sum_{i=1}^q \theta_i a_{t-i} \quad (1)$$

The three basic parameters, namely p , d and q involved in (1) represent respectively, the amount of autoregression, level of systematic change over time (trend) and moving average part. In the modeling process these three parameters are estimated in an iterative way using three stages, viz. model identification, parameter estimation and diagnostic checking until the most suitable model is found (Chen, 2008, Bisgaard and Kulachi, 2011).

The basis for any time series analysis is stationary time series (Chen, 2008, Bisgaard and Kulachi, 2011). A stationary time series has constant mean, constant variance and constant autocorrelation structure. Therefore, first step in

developing an ARIMA model is to test if the series is stationary. Three ways were adopted to ascertain stationarity. These were as follows:

- i. Examining the plot of the raw data. To be stationary, the plot of the series should show constant location and scale.
- ii. Observing the autocorrelation function (ACF), partial autocorrelation function (PACF) and the resulting correlograms.

The ACF at lag k , denoted by ρ_k , is defined as:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\text{Auto covariance at lag } k}{\text{Varinace of the time series}} \quad (2)$$

Where $k \leq T/4$, T = Number of observations. In the present study, we have 64 observations. If ACF does not damped out within $64/4=16$ lags, the process is likely to be non- stationary (Bisgaard and Kulahci, 2011).

- iii. To provide further evidence for the nonstationary time series, we conducted Augmented Dickey-Fuller (ADF) Unit Root Test (Gujarati and Sangeetha, 2011).

If the time series is stationary, then the assumption of constant mean and homogeneity variance are met. However, if the pattern presents a trend, the method of differencing advocated by Box-Jenkins can be used to remove the linear or curvilinear trend. The first order of differencing ($d = 1$) is designed to remove the linear trend while the second order of differencing ($d=2$) is used to remove the curvilinear trend (Chen, 2008).

The variability (if any) present in a process may be stabilized by employing logarithmic transformation before first differencing (Negron [14]). A rough graphical check for the right transformation is the range-mean plot, which is produced by dividing the time series into smaller segments and plotting the range versus the average of each segment on a scatter plot. If the plot indicates linear relationship between average and range, then log transformation is appropriate (Bisgaard and Kulahci [6], pp. 114-116).

Once the stationarity of the nonstationary time series for female examinees is achieved, next step is to identify the order of ARIMA (p, d, q) model. The primary tools in identification are the Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF) and the resulting correlograms. After identifying the appropriate p and q values, the parameters included in the model are estimated. To choose the best model we used the Akaike's Information Criteria (AIC) (Deb Roy and Das, 2012). We choose the model that has

minimum AIC value. For a sample size of n observations, AIC is given by (Bisgaard and Kulahci, 2011):

$$AIC = -2\text{Log}(\text{Maximized Likelihood}) + 2r \approx n\text{Log}(\hat{\sigma}_a^2) + 2r \quad (3)$$

Where $\hat{\sigma}_a^2$ is the maximum likelihood estimate of the residual variance σ_a^2 , r is the number of parameters estimated in the model including a possible constant.

We also applied forecast accuracy criteria, mean absolute percentage error (MAPE), root mean square error (RMSE) and mean absolute error (MAE) to decide the better model. The RMSE, MAPE and MAE are defined as:

$$\text{Root mean square error, } RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

Mean absolute percentage error,

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (5)$$

$$\text{Mean absolute error, } MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

where y_i is the observed value, \hat{y}_i is the predicted value and n is the number of predicted values. The smaller the value of $RMSE$, $MAPE$ and MAE , the better the model is (Chen, 2008, Shitan and Lerd, 2014).

For further check to ascertain how well the estimated model fits the data, we conducted the usual residual diagnostic test. If the residuals estimated from the selected model are white noise, we can accept the particular fit (Gujarati and Sangeetha, 2011). To confirm, we further used Ljung-Box statistic to test for non-zero autocorrelations in the residuals at lags 1-16. A white noise or purely random process has zero mean and constant variance σ^2 . The Ljung-Box (LB) statistic is defined as,

$$LB = n(n+2) \sum_{k=1}^m \left(\frac{\rho^2}{n-k} \right) \sim \chi_m^2 \quad (7)$$

If observed χ_m^2 is less than expected χ_m^2 then there is no evidence of white noise at lags 1-16 (Gujarati and Sangeetha, 2011). The R software was used for fitting ARIMA model in the study and also for outliers detection using the *tsoutliers* package.

5. Results and Discussion

5.1 Female examinees

In this section we consider the analysis of the female examinees in Matriculation/ HSLC Examination in Assam from 1951 to 2015.

5.1.1 Test for Stationarity

The time series plot of female examinees in Matriculation/HSLC Examination in Assam from 1951 to 2015 presented in Figure 1 (a) reveals that the data is non-stationary i.e. a chain of rapid growth, sudden decline, sudden growth and uprising trends. The non-stationary pattern was further confirmed by observing the ACF and PACF plots in Figure 1 (b) and Figure 1 (c). These plots depict strong positive autocorrelation. It is observed that the ACF for raw data does not die out even for large lags. Therefore, the time series is nonstationary (Bisgaard and Kulahci, 2011, Gujarati and Sangeetha, 2011). Autocorrelation at lag 2 and above are merely due to the propagation of the autocorrelation at lag 1. This is confirmed by Partial Autocorrelation Function (PACF) plot. The PACF plot has a significant spike at lag 1, meaning that all the higher order autocorrelation are effectively explained by the lag-1 autocorrelation (people.duke.edu/~rnau/411arim3.htm). The non-stationarity of the time series is confirmed by using Augmented Dickey-Fuller (ADF) test. With a Dickey-Fuller test statistic = -0.4257 and p -value = 0.9823, we accept that the time series for female examinees in the examination is nonstationary. The non-stationarity in mean was corrected by differencing the data.

Before differencing we have checked for right log transformation to obtain homogeneous variability, by dividing the time series into small segments of five years, computed the five years' range and average and plotted as scatter plot in Figure 2(a) for female examinees appeared in the examination. Presence of relationship is observed from the Figure 2(a) between averages and ranges of segments in the time series. Therefore, log transformation is done and presented the respective graph in Figure 2(b). A look at the Figure 2 (b) and Figure 1(a) reveals that the variability is considerably reduced upon applying the transformation.

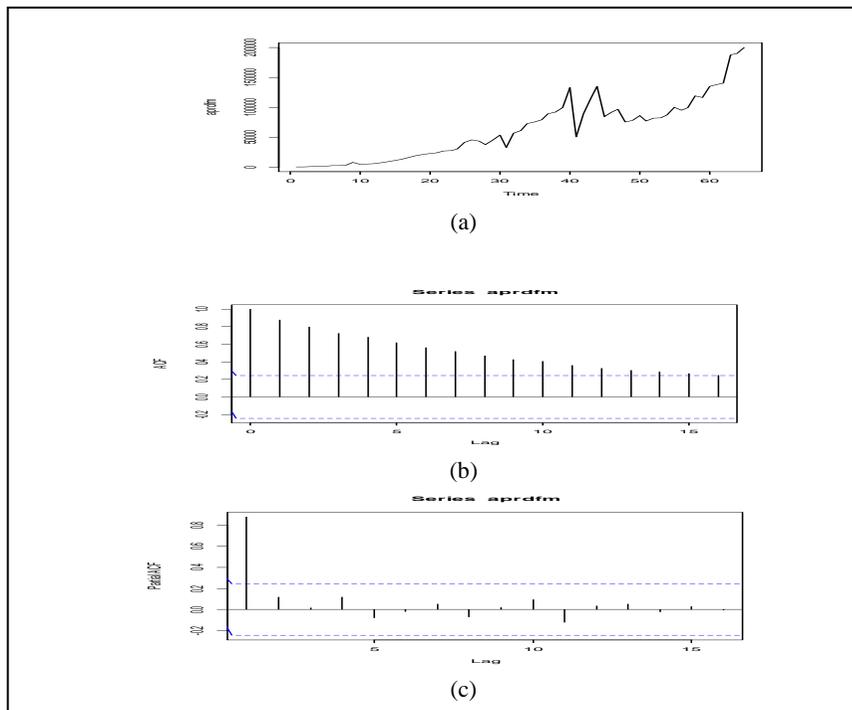


Figure. 1 (a) Time Series plot (b) ACF plots (c) PACF plots of Female Examinees in Matriculation/ HSLC examination in Assam (1951-2014)

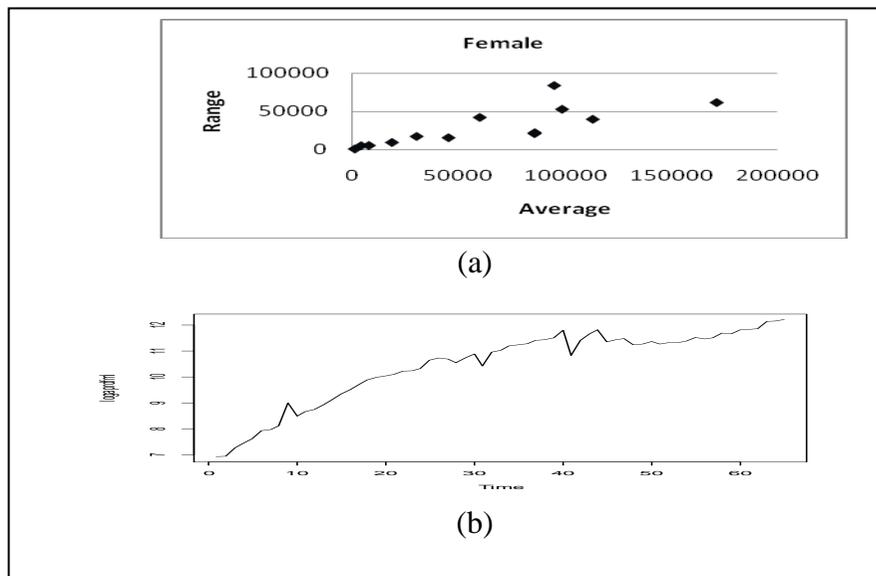


Figure. 2 (a) Range-Mean plot (b) Log Transformed series of Female Examinees in Matriculation/ HSLC Examination in Assam (1951-2015).

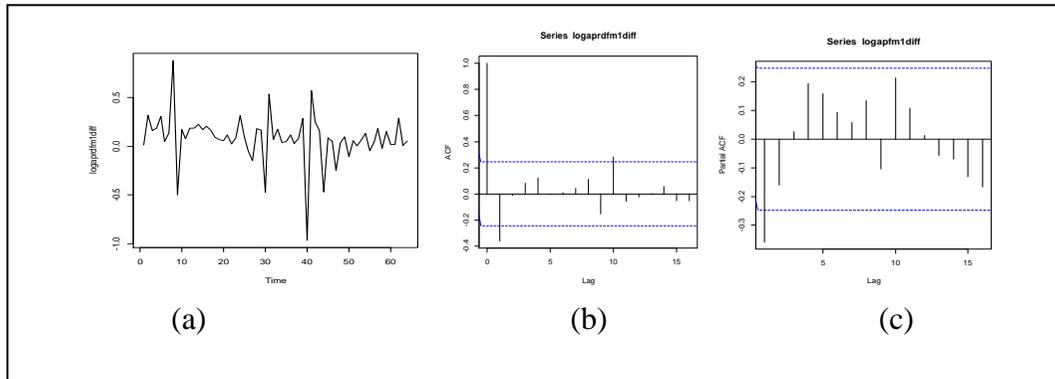


Figure. 3 Plot of log-transformed first differenced (a) Time Series (b) ACF (c) PACF of Female Examinees in Matriculation/ HSLC Examination in Assam (1951-2015)

Figure 3 (a) represents log-transformed first differenced time series plot, Figure 3 (b) and 3 (c) represent the ACF and PACF plots of log-transformed first differenced series. The figures suggest that, as a result of log transformation and first differencing, non-stationarity is considerably reduced. For confirmation we apply ADF test to the log-transformed first differenced series. It is found that Dickey-Fuller test statistic = - 4.2876 with p -value 0.01. From the ACF and PACF plots and ADF test we confirmed that the series is now stationary.

5.1.2 Model Identification

For identifying the ARIMA (p, d, q) model, we examine the ACF and PACF plots of stationary time series of female examinees in Matriculation/ HSLC Examination in Figure 3(b) and Figure 3(c). Here we differenced the series only once, so $d = 1$. For identifying the order of autoregressive component p we observed the PACF plot in Figure 3(c). It is observed that, there is one large autocorrelation at lag 1. All other autocorrelations cut off after lag 1. Therefore PACF plot suggests that AR (1) model could be accurate in the present case.

To get idea about the order of moving average component, we observed the ACF plot in Figure 3(b). Here we noticed, there is a significant autocorrelation at lag 1 after which all other autocorrelation drop near to zero. Therefore, ACF plot suggests that MA (1) model may be accurate.

Now we have identified p, d, q , so our potential models may be:

1. ARIMA (1, 1, 0) as PACF is zero after lag 1.
2. ARIMA (0, 1, 1) as ACF is zero after lag 1,
3. ARIMA (1, 1, 1) a combination model of 1 and 2.

We tried all the models and from the point of view of parsimony and forecast accuracy we selected the model containing lowest value of AIC, MAPE, RMSE and MAE. The results are presented in Table 1.

Table. 1 The AIC, RMSE, MAE and MAPE values for the Time Series of Female Examinees in Matriculation/ HSLC Examination in Assam (1951-2014)

Model	AIC	RMSE	MAE	MAPE (%)
ARIMA (1,1,0)	9.28	0.2500	0.1792	1.7920
ARIMA (0,1,1)	9.81	0.2512	0.1798	1.7926
ARIMA (1,1,1)	11.24	0.2500	0.1793	1.7852

In the present study, all models are seemed to be adequate. However, after residual analysis we preferred ARIMA (1, 1, 0) model over ARIMA (0, 1, 1) and ARIMA (1, 1, 1) models.

5.1.3 Model Estimation

After identification of the model, the parameters are estimated. In Table 2 the summary results of the fitted models are incorporated.

Table. 2 Estimated coefficients for ARIMA (1, 1, 0) Model (Female)

Variable	Coefficient	Standard Error	σ^2	Log-likelihood	<i>p</i> -value
AR(1) ϕ	-0.2160	0.1210	0.0635	-2.64	0.083

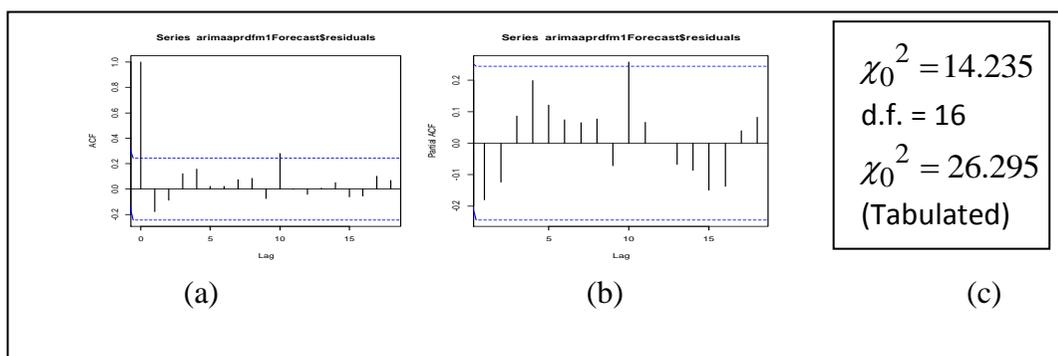


Figure. 4(a) ACF Plot (b) PACF Plot (c) Chi Square value of Ljung-Box test of residuals (Diagnostic Tests)

5.1.4 Diagnostic Checking

For further check how well the ARIMA (1, 1, 0) model fits the data, we conducted usual residual diagnostic checks. Figure 4(a) to Figure 4(c)

represent the ACF, PACF and Ljung-Box Chi square value of residuals after fitting ARIMA (1, 1, 0) model to the time series of female examinees under study. It is observed that none of these results indicate any significant autocorrelation.

The estimated ARIMA (1, 1, 0) model was applied to forecast the future female examinees in the examination. The observed data can be treated as realization from an infinite population of such time series that could have been generated by the stochastic process (Bisgaard and Kulahci, 2011). So, exact prediction of future values is not possible. Therefore, a prediction interval and the probability with which the future observation will lie within the interval can be provided (Bisgaard and Kulahci, 2011). If $\sigma_e^2(l)$ be the variance of prediction errors, then 95% prediction interval is given by

$$\hat{z}_t \pm 1.96\sigma_e(l) \quad (8)$$

We attempted to forecast the female examinees in the examination for five periods ahead, i.e., for 2016, 2017, 2018, 2019 and 2020. The results are presented in Table 3. The comparison graph of forecasts, the 95% prediction intervals and actual observations are plotted in Figure 5. From visual observation, it is clear that the chosen model is reasonably good as the predicted series is very close to the actual series and within the 95% prediction limits.

Mean while, the process of the HSLC Examination 2016 which started on 18th February has ended with declaration of the results on 31st June 2016. The number of female examinees in the examination is 199763. It is important to mention here that the number of female examinees this time was 199763, which is very close to the forecasted value of 198154 by our model. That is the relative error is less than 1%.

Table. 3 Forecast with 95% prediction intervals for Female Examinees with ARIMA (1, 1, 0) Model

Year	Forecast	95% LPL	95% UPL
2016	198154	105787	371172
2017	198051	93571	419190
2018	198073	84465	464488
2019	198069	75062	522646
2020	198069	68882	569549

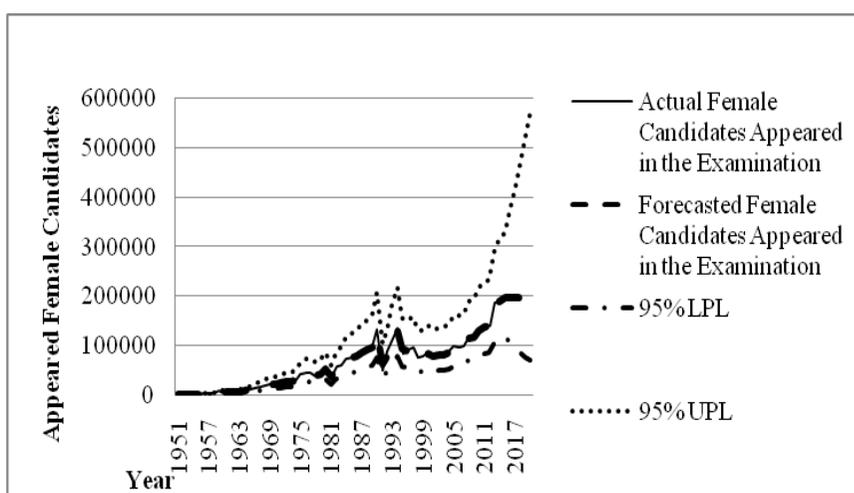


Figure. 5 Comparative graph of Actual vs. Forecasted values of Female Examinees in Matriculation/ HSLC Examination in Assam (1951-2019) with 95% prediction levels

5.2 A comparison with Male examinees

Though the primary objective of the present study is to propose a model of female examinees in Matriculation/HSLC examination in Assam from 1951 to 2015, in this section we consider the male examinees. Figure 6 depicts the trend of both female and male examinees from where it is observed that, the appeared rates of both male and female examinees have increased over time and in the course of time number of female examinees surpassed the number of male examinees.

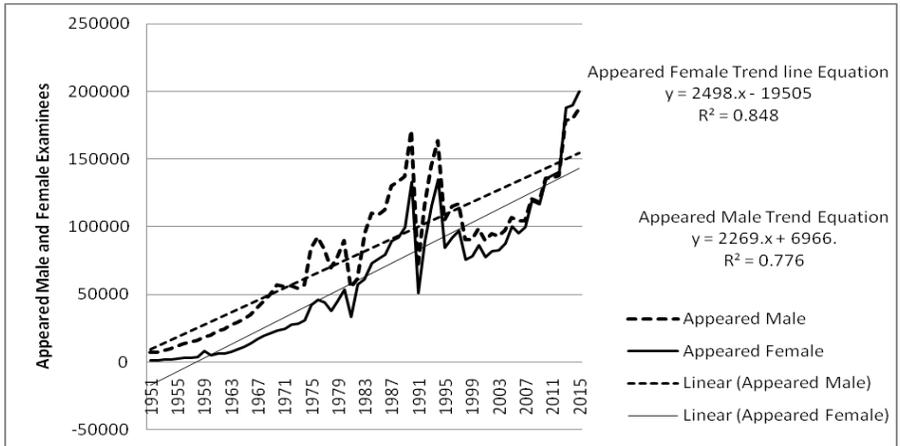


Figure. 6 Graph depicting trend of both Male and Female Examinees for 1951-2015

Here like in the case of female examinees we carried out the same sequence of test of stationarity, log transformation, model identification, model estimation, diagnostic checking and finally forecasting. The summary results are presented in Figure 7 to Figure 12 and Table 4 to Table 6 leads to ARIMA (2, 1, 0) as the most suitable model for forecasting male examinees in the examination as opposed to ARIMA (1, 1, 0) for females.

5.2.1 Test for Stationarity

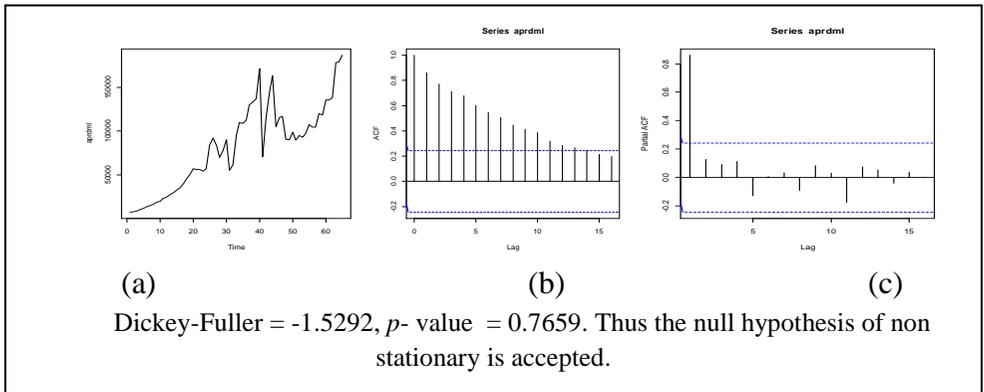


Figure. 7(a) Time Series plot (b) ACF plot (c) PACF plot and Dickey-Fuller test result of Male Examinees in Matriculation/ HSLC Examination in Assam (1951-2015)

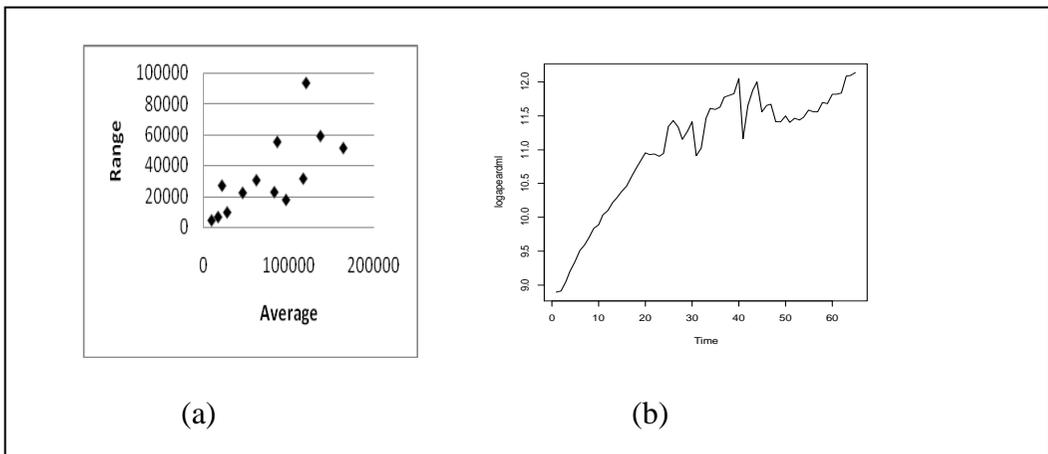
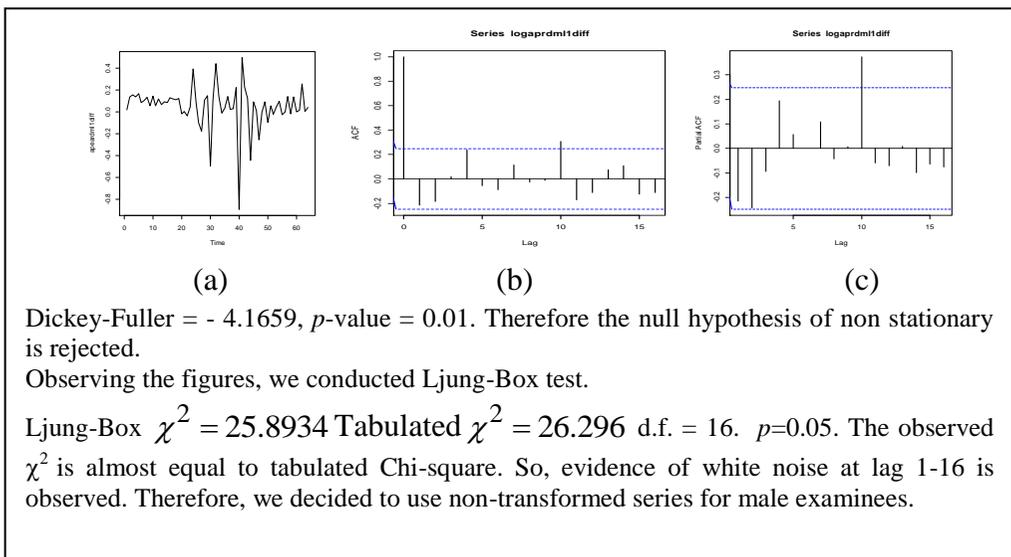


Figure. 8(a) Range-Mean plot (b) log transformed series of Male Examinees in Matriculation/HSLC Examination in Assam (1951-2015)



Dickey-Fuller = - 4.1659, p -value = 0.01. Therefore the null hypothesis of non stationary is rejected.

Observing the figures, we conducted Ljung-Box test.

Ljung-Box $\chi^2 = 25.8934$ Tabulated $\chi^2 = 26.296$ d.f. = 16. $p=0.05$. The observed χ^2 is almost equal to tabulated Chi-square. So, evidence of white noise at lag 1-16 is observed. Therefore, we decided to use non-transformed series for male examinees.

Figure: 9 (a) Time Series (b) ACF (c) PACF Plots , Dickey-Fuller Test, Ljung-Box test of log-transformed First Differenced Time Series (Male Examinees in Matriculation/HSLC Examination 1951-2015)

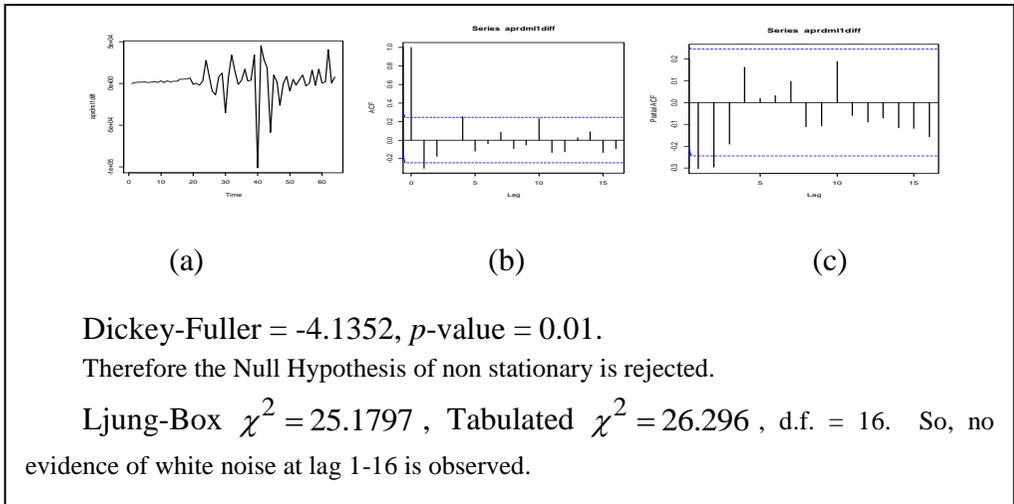


Figure. 10 (a) Time Series plot (b) ACF plot (c) PACF plot, Dickey-Fuller test, Ljung-Box test of First Differenced Time Series (Male Examinees in Matriculation/ HSLC Examination 1951-2015)

5.2.2 Model Identification [From Figure 10 (b) and Figure 10 (c)]

Table. 4 Results of Model Identification for Male Examinees

(1)	ARIMA (2, 1, 0) as the PACF is zero after lag 2.			
(2)	ARIMA (0, 1, 1) as the ACF is zero after lag 1			
(3)	ARIMA (2, 1, 1) combination of (1) and (2).			
Models With AIC, RMSE, MAE and MAPE (%)				
Model	AIC	RMSE	MAE	MAPE (%)
ARIMA(2,1,0)	1446.07	14.6186	18569.59	11612.30
ARIMA(0,1,1)	1446.16	14.8634	18750.03	11765.70
ARIMA(2,1,1)	1448.57	14.8947	18510.14	11770.15

Table. 6 Forecast with 95% Prediction Intervals for Male Examinees with ARIMA (2, 1, 0)

Year	Forecasted	95% LPL	95% UPL
2016	184076	147397	220755
2017	183160	139272	227047
2018	184161	136148	232173
2019	184041	130308	237774
2020	183838	125127	242548

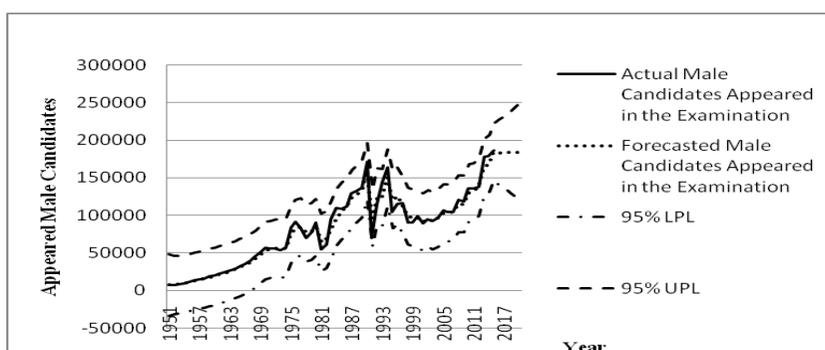


Figure. 12 Comparative graphs of Actual Vs Forecasted Male Examinees in Matriculation / HSLC Examination in Assam (1951-2020) (With 95% upper and lower prediction limits)

6. Detection and Removal of Outliers

The time series data of both female and male examinees of Matriculation/HSLC Examination considered in this study show sudden changes. Such changes which may be attributed to presence exogenous or outlier effects may alter the dynamics of the series either permanently or transitory. As such detecting outliers is important as they may have impact on the selection of model, the estimation of parameters and consequently on forecasts (Lopez-de-Lacalle, 2016). In this section we look for potential outliers using R-Package “tsoutliers” (Lopez-de-Lacalle, 2016). For both female and male data set ten outliers are detected as summarized in Table 7.

Table. 7 Outliers detected for Female Examinees

Sl. No.	Type of outliers	Time point	tstat
1	AO	30	3.659
2	LS	25	4.177
3	LS	33	4.410
4	TC	43	4.244
5	TC	31	-5.717
6	AO	40	10.595
7	TC	41	-11.830
8	AO	44	7.945
9	LS	45	-6.060
10	LS	63	5.854

Table. 8 Outliers detected for Male Examinees

Sl. No.	Type of outliers	Time point	tstat
1	AO	40	9.610
2	LS	33	4.965
3	TC	42	4.387
4	TC	43	4.003
5	LS	25	4.236
6	TC	31	-5.615
7	AO	41	-12.331
8	AO	44	7.909
9	LS	45	-6.076
10	LS	63	6.297

[Here AO = Additive outlier, IO =Innovative outlier, LS = Level shift outliers and TC = Temporary change outliers. To determine the significance of each type of outlier the default critical value to 3.5. (see page 11of the package ‘tsoutliers’, July 7, 2016).]

Table. 9 Forecast with 95% Prediction Intervals after detection and Interpolation of outliers Female Examinees [ARIMA (1, 1 ,0) model]

Year	Forecasted (Female)	95% LPL	95% UPL
2016	198622	122520	321995
2017	198777	104557	377902
2018	198759	91703	430792
2019	198761	82032	481586
2020	198761	74304	531682

In order to overcome the possible effect of these outliers usually in the case of non-seasonal time series like ours, outliers are replaced by linear interpolation. (Hyndman, 2014). Therefore, we have first subjected the time series to linear interpolation to get rid of the outliers and carried out same sequence of test of stationary, log transformation, model identification, model estimation, diagnostic checking and finally forecasting for two new series and found that the same ARIMA(1,1,0) and ARIMA(2,1,0) models for both female and male examinees respectively. The forecasted values after the interpolation of outliers for female and male presented in Table 8 and Table 9 are almost the same as they were before outlier interpolation (see Table 3 and Table 6).

Table. 10 Forecast with 95% Prediction Intervals after detection and Interpolation of outliers Male Examinees [ARIMA (2, 1, 0) model]

Year	Forecasted (Female)	95% LPL	95% UPL
2016	184073	147405	220740
2017	183155	139287	227022
2018	184159	136177	232142
2019	184039	130339	237738
2020	183835	125162	142507

7. Concluding Remarks

The numbers of female and male examinees in Matriculation/ HSLC Examination in Assam for the period 1951 to 2015 have been studied using Box-Jenkins ARIMA modelling technique. The ARIMA (1, 1, 0) and ARIMA (2, 1, 0) models were identified as the best model for female and male examinees respectively. From the estimation and diagnostic results it is evident that the models are adequately fitted to the data under study, which is further confirmed by carrying out the residual analysis. The selected models ARIMA (1, 1, 0) and ARIMA (2, 1, 0) are then used to predict five years female (male) examinees for 2015 to 2020 in the Examination.

We have detected outliers in the both the time series. It is verified that even after the interpolation of the detected outliers our analysis led to the selection same models as the ones selected based on the original data for both female and male examinees. It may therefore be concluded that for both the time series data considered in this analysis are robust and not much affected by the presence of the outliers.

It is pertinent here to mention that, various factors may determine the number of examinees. Some of which may be central and state financial aids, increase of number of secondary schools, infrastructure facilities, quality of instruction etc. Though the historical data on such factors are not considered in the present study, yet it can help the policy builders to adopt appropriate strategies for the progress and empowerment of female in the state. The adequacy of our proposed models can also be seen from the very low relative forecasting error for the current year examinees.

References

1. Aggarwal, J. C. (2009): Development and Planning of Modern Education, 9th Edition, Vikas Publishing House Pvt. Ltd, New Delhi.
2. Asian Development Bank, UNDP and UNESCAP (2006): *Pursuing Gender Equality through the Millennium Development Goals in Asia and the Pacific*. Publication Stock No. 050406. Asian Development Bank, Manila.http://onesearch.slq.qld.gov.au/SLQ:SLQ_PCI_EBSCO:slq_alm_a21135533650002061
3. Bisgaard, S. and Kulahci, M. (2011): *Time Series Analysis and Forecasting by Example*. Wiley, A John Wiley & Sons, INC., Publication.
4. Board of Secondary Education, Assam (2010): *Results of High School Leaving Certificate Examination & Assam High Madrasa Examination*.
5. Box, G. E. P., Jenkins, G. M. and Reinsel, G. C.: *Time Series Analysis, Forecasting and Control*. 3rd Edition, Pearson Education, 2003.
6. Chen, C-K. (2008): *An Integrated Enrollment Forecast Model*. Association for Institutional Research. IR Applications, Number 15, 1-18.
7. Decision 411 Forecasting: *Identifying the numbers of AR or MA terms*, people.duke.edu/~rnau/411arima3.htm
8. Deb Roy, T. and Das K. K. (2012). *Time Series Analysis of Dibrugarh Air Temperature*. Journal of Atmospheric and Earth Environment. British Academic Journals 1(1), 30-34.
9. Document of the World Bank (2009) *Secondary Education in India Universalizing Opportunity*. Human Development Unit South Asia Region, 2009. World Bank, Washington, DC: [http:// documents. Worldbank.org/ curated/en/262201468285343550/Secondary-education-in-India-universalizing- opportunity](http://documents.worldbank.org/curated/en/262201468285343550/Secondary-education-in-India-universalizing-opportunity)
10. Gauhati University Examination Results (1948-1968): *Compiled in the Statistical Unit of the Gauhati University under supervision of the Statisticians*, Gauhati University.

11. Gujarati, N. D. and Sangeetha (2007). *Basic Econometrics*, Fourth Edition. Tata McGraw Hill Education Pvt. Ltd., New Delhi.
12. Hyndman, R. (2014). Stats.stockexchange.com/questions/69874
13. Lopez-de-Lacalle, J. (2016). Package ‘tsoutliers’ tsoutliers R package for detection of outliers in time series, <http://cran.r-project.org/web/packages/tsoutliers/tsoutliers.pdf>
14. Negron, A. (2013): *ARIMA model in a Nutshell*, Foundation of open Access Statistics alton.github.io/r/2013/05/
15. Rashtriya Madhyamic Shiksha Abhiyan (RMSA)-India: www.indg.in, April 3, 2012.
16. Shitan, M., Karmoker, P. K and Lerd, N. Y. (2014): *Time Series Modeling and Forecasting of Ampang Line Passenger Ridership in Malaysia*. Pakistan Journal of Statistics, 30(3), 385-396.
17. The Assam Tribune, Guwahati, Wednesday, November 11, 2009: Editorial, *Class-X public examination-a burden or boon?*
18. The Times of India, Guwahati, December 3, 2013: *More girls than boys enroll at secondary level, RMSA.*