Forecasting of Female Labor Force Participation Rate data with missing values imputation, Sri Lanka

NN Ranasinghe[1] and RAB Abeygunawardana[1]

[1]*Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka*

Abstract: *Female Labour Force Participation Rate (Female LFPR) is defined as the proportion of the female labor force to the total working-age population. This study was based on the female LFPR quarterly data published by the Department of Census and Statistics, Sri Lanka from 2004 to 2021. However, it was found that data for eight quarters are missing in the above period. The main objective of this study is to forecast female LFPR using ARIMA models by imputing the missing values. In the first part of the analysis, missing values were imputed using nine imputation algorithms available in "imputeTS" package in R software. Missing values were generated under four missing rates and thirty random seeds. By comparing MAPE and RMSE plots the Exponential Weighted Moving Average (EWMA) method was found to be the best imputation method. In the second part of the analysis, female LFPR were forecasted using ARIMA models. In this analysis, the data were divided into two parts as training and test data. In the training data set, trend, seasonal and random components were identified using the "decompose()" function in R software. Furthermore, functions "arima()" and "auto.arima()" in library "forecast" in R software were used to fit ARIMA models. It was found that ARIMA(1,1,1) model without drift was the best model to forecast the female LFPR which has the minimum AIC value. Errors for the fitted values were calculated using the test data. Female LFPR for the next ten quarters was forecasted using the ARIMA(1,1,1) model. Results showed a small increment in female LFPR at the end of 2022.*

Keywords: *LFPR, ARIMA models, Imputation*

1. Introduction

The labor force includes both number of people employed and unemployed. The labor force participation rate (LFPR) is a key measure of labor force analysis. Analyzing gender-wise LFPR is important because the contribution of males and females are not the same in the labor force. As a developing country, getting more contributions to the labor force by females is important because it helps to achieve economic stability and improve social well beings. Further, women as a mother play a major role in the family. When they are employed, the standard of living of their families is improved and their lives become more comfortable economically. Therefore, it is important to identify changing patterns of female LFPR data and predict future values to make policies to increase the female contribution to the labor force.

This study was based on female labor force participation rate quarterly data which are published by the Department of Census and Statistics (DCS) from 2004 to 2021. Female LFPR is defined as the proportion of the female labor force to the total working-age population. Working-age people are defined as those who are 15 years old or older after 2013. Before 2013, this was defined as a person who was 10 years old or older. The labor force survey was started by DCS in 1990. But DCS was unable to conduct a labor force survey (LFS) in a few quarters due to several reasons. LFS was not conducted in the second quarter of 2001 due to the heavy workload of the Census of Population and Housing of 2001. Again, due to the Tsunami, LFS was not implemented quarterly as planned in

2005. LFS was not conducted in the 4th quarter of 2011 and 1st quarter of 2012 also due to the Census of Population and Housing in 2012. Since 2013, the survey has been done in all four quarters of each year, covering the entire country.

The objective of this study is to forecast female LFPR using ARIMA models with missing values imputations to see how the unemployment rate will look in the future.

2. Methodology

This study was done in two parts. The first part is missing values imputation and the second part is model fitting and forecasting using ARIMA models.

A. About Data

Female LFPR data which were published by DCS, Sri Lanka, from the 1st quarter of 2004 to the 2nd quarter of 2021, were used for the analysis (70 data points). Data in all quarters of 2005, the first two quarters of 2006, 4th quarter of 2011, and 1st quarter of 2012 were missing data. Hence 8 data points were missing in the considered period.

B. Missing values imputation

Nine different imputation methods in package "*imputeTs*" in R software were compared to select the best imputation algorithm for missing values imputation. Part of the data set without really missing values was selected as complete series for missing values imputations analysis.

Female LFPR data from the 2nd quarter of 2012 to the 2nd quarter of 2021 (36 data points) was considered as the complete series. Missing values were randomly generated using the Bernoulli distribution under four missing rates such as 0.1,0.25,0.5 and 0.8. The success probability of the Bernoulli distribution equals to missing rate. When the generated missing value equals 1, the corresponding value in the time series is replaced by NA (Not Available) and from now on is considered to be missing. Since the results of the imputation algorithms can be influenced by the pattern of missing data, the function generates the missing data by running with 30 different random seeds, to randomize the results. Results were based on experiments for 30 random seeds, 4 levels of messiness, implementing 9 imputation algorithms, that is 1080 runs for this data set. Imputation algorithms that were used for missing values imputations are shown in table 1.

Table 3: Overview of imputed algorithms

Function	Option	Description
na.kalman	StructTS	Imputation by Structural Model & Kalman Smoothing
	auto.arima	imputation by ARIMA State Space Representation & Kalman Smoothing.
na.interpolation	linear	Imputation by Linear Interpolation
	spline	Imputation by Spline Interpolation
	stine	Imputation by Stineman Interpolation
na.ma	simple	Missing Value Imputation by Simple Moving Average
	linear	Missing Value Imputation by Linear Weighted Moving Average
	exponential	Missing Value Imputation by Exponential Weighted Moving Average
na.mean	mean	Missing Value Imputation by Mean Value

1)Evaluating Imputation Accuracy; Two error metrics of mean root square error (MRSE) and mean absolute percentage error (MAPE), were used to measure the effectiveness of the imputation algorithms. Considering MRSE and MAPE values best imputation algorithm was selected for each variable.

Define y_i as the ith observation in the complete series. For the realization of the time series for a specific random seed and rate of missing data, $\widehat{y_\iota}$ is the imputed value and n is the number of missing values. The equation then yields MRSE.

$$\text{MRSE}(\widehat{y_\iota},\text{y}_\text{t}) = \sqrt{\frac{\sum_{t=1}^{n}(\widehat{y_t}-y_t)^2}{n}}$$

and MAPE is given by equation

$$\text{MRSE}(\widehat{y_\iota},\text{y}_\text{t}) = \frac{\sum_{t=1}^{n}\left|\frac{\widehat{y_t}-y_t}{y_t}\right|}{n} \times 100\%$$

C. Training and test data

Unemployment rate data was divided into two parts, test and training data. The training data contains quarterly female LFPR data from the 1st quarter of 2004 to the 2nd quarter of 2021(66 data points). Female LFPR data from the 3rd quarter of 2020 to the 2nd quarter of 2021 was considered as test data (4 data points). Training data was decomposed into three components such as trend, seasonal and random using the function "decompose ()". Changing behavior of these components was examined using these decomposition plots.

D. ARIMA (p, d, q) model fitting

Stationarity was tested using the function "*adf. test ()*" in R which is relevant to Augmented Dickey-Fuller (ADF) test, where the null hypothesis indicates that the series is non-stationary. The first-order deference series was stationary. ACF and PACF plots of 1st difference series of female LFPR were used to initiate the order of MA terms and order of AR terms respectively. Then different order ARIMA (p, d, q) models were fitted using functions "*auto.arima()*" and "*arima ()*". Then AIC values and the significance of coefficients of all fitted models were compared. Then, the model with min AIC and significant coefficients at 5% was selected as the best model.

E. Female LFPR forecasting

Female LFPR was forecasted from the 3rd quarter of 2020 to the 4th quarter of 2022, using the selected ARIMA (p, d, q) model. MAPE and RMSE of predicted values were calculated using test data relevant to the 3rd quarter of 2020 to the 2nd quarter of 2021.

F. Adequacy of the fitted model

Model adequacy was measured using residual analysis. Function "*box. test ()*" which is relevant to the Ljung-Box test was used for examining the null hypothesis of independence in given residuals. The Shapiro-Wilk's test or Shapiro test is used to test the normality of residuals. To perform the Shapiro-Wilk test, the function of "*shapiro. test ()*" in R was applied.

3. Results

Missing values imputation results are as follows:

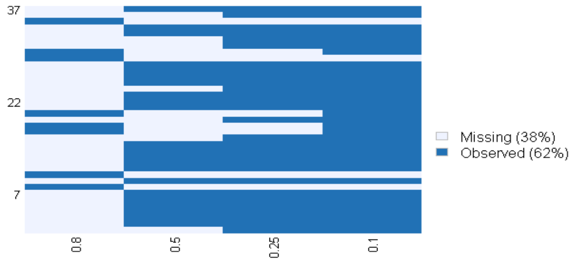

Figure 1: Missing map of generated missing data

The missing map in figure 1 was produced by "*imputeTs*" package in R (Elissavet, 2017). Missing data patterns for different levels of missingness were displayed.

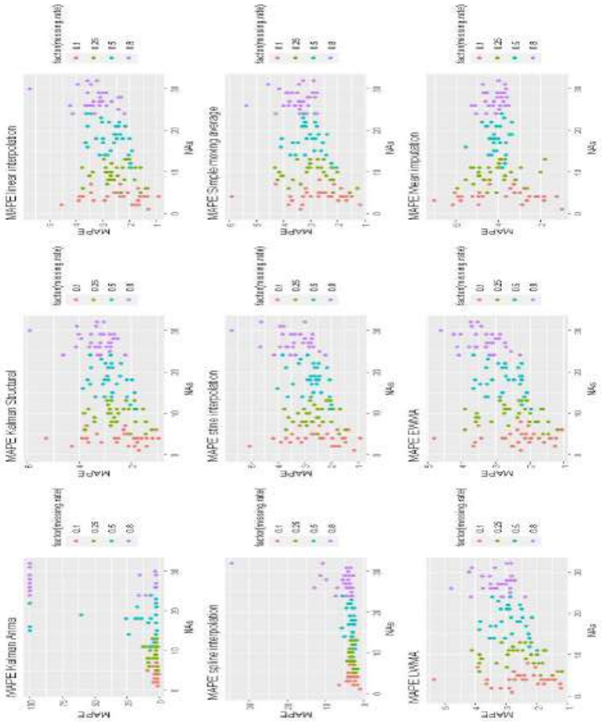

Figure 2: Plots of MAPE values

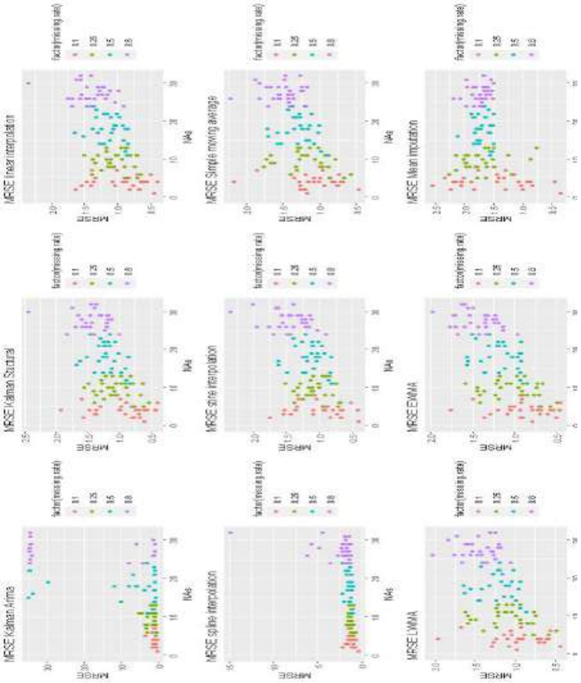

Figure 3: Plots of RMSE

The imputation method described in table 1 was used to impute generated missing values. Figure 2 represents plots of MAPE computed by different imputation methods at different missing rates. The point of the plot relates to MAPE values of random seeds and colors indicating each missing rate. Figure 3 represents plots of RMSE values, represents plots of RMSE computed by different imputation methods at different missing rates. Kalman arima, spline interpolation shows relatively high error values. Especially, RMSE imputed by the Kalman arima method shows high error values at the level of missing rates equal to 0.5 and 0.8. The mean imputation method shows relatively high RMSE and MAPE values at lower missing rates than high missing rates. RMSE distribution from Kalman structural and linear interpolation methods show a similar pattern. In all other methods, RMSE values show an increasing pattern when the missing rate is increased. Considering RMSE plots exponential weighted moving average (EWMA) method was selected as a more suitable imputation method. EWMA method shows minimum error distribution in MAPE plots also. Therefore, by considering both MAPE and RMSE distributions, the EWMA method was selected as the best method for missing values imputation of female LFPR. Imputed missing values using EWMA) the method is illustrated in table 2.

Table 4: Imputed missing values by EWMA method

Year	Quarter	Imputed Female LFPR
2005	1	35.1800
2005	2	35.18571
2005	3	36.20000
2005	4	38.30000
2006	1	38.90000
2006	2	38.72667
2011	4	34.05455
2012	1	33.30455

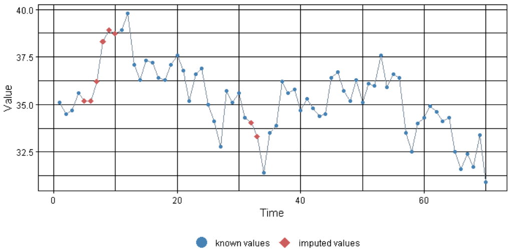

Figure 4: Plot of Female LFPR with imputed missing values

Figure 4 visualizes a time series plot of female LFPR from the 1st quarter of 2004 to the 2nd quarter of 2021. Female LFPR values are varying from 39.8% to 30.9% where the highest value is in the 4th quarter of 2006 and the lowest value is in the 2nd quarter of 2021. As an overall picture, the long-term trend of female LFPR is a decline. The seasonal components of each quarter represent in table 3. The second and 3rd quarters represent negative seasonality. The highest seasonal index exists in the 2nd quarter.

Table 5: Seasonal components of Female LFPR

Quarter	Seasonal index
1	0.17215338
2	-0.45891940
3	-0.04241245
4	0.32917846

ARIMA model fitting and forecasting results are as follows:

ADF test results for the 1st-order difference series are shown in table 4. Since the p-value of this test is less than 0.05, the null hypothesis was rejected at a 5% level of significance by concluding that the series is stationary.

Table 6: ADF test for 1st order difference series of Female LFPR

Augmented Dickey-Fuller Test	
Dickey-Fuller	-4.3289
Lag order	4
p-value	0.01

ACF and PACF plots of 1st order difference series were used to identify the significant number of MA and AR terms respectively.

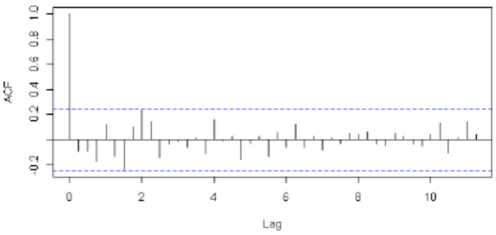

Figure5: ACF plot of 1st order difference series of female LFPR

ACF plot of 1st order difference series of female LFPR represents in figure 5. There is a cut-off at lag zero. There is a specific pattern in autocorrelation coefficients. All autocorrelation coefficients are not significantly different from zero because all autocorrelation coefficients lie within the confidence band except autocorrelation coefficients at lag zero. Therefore, the order of MA terms was initiated from zero.

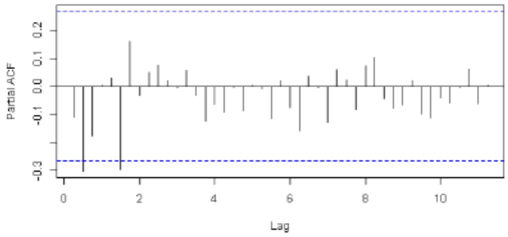

Figure 6: PACF plot of 1st order difference series of female LFPR

PACF plot of 1st order difference series is shown in figure 6. Partial autocorrelation at lags 2 and 6 is significant. Order of AR term was initiated from 2. Different ARIMA models were fitted by initiating from ARIMA (6,1,0) model. The function of "*auto. arima ()*" suggested ARIMA (0,1,0) without drift model (random walk mode) as the best. Considering the minimum AIC value and significance of coefficients ARIMA (1,1,1) without drift model was selected as the best model. Function "*coeftest ()*" in the library "*lmtest*" was used to test the significance of coefficients. The estimated coefficients of the fitted model represent in table 5.

Table 7: ARIMA(1,1,1) without drift model for female LFPR

	AR1	MA1
Coefficients	0.7615	-0.9515
Standard Error	0.1236	0.0731
Sigma^2		1.298
log-likelihood		-100.02
AIC		206.03
AICc		206.42
BIC		212.55

Table 6 visualizes z values and corresponding p values for coefficients of fitted models. Since p values are less than 0.05 it was concluded that coefficients are significant.

Table 8: Z test of coefficients

	Estimate	Std. error	Z value	Pr(>|Z|)
AR(1)	0.761480	0.123568	6.1625	7.163e-10
MA(1)	-0.951503	0.073076	-13.0207	2.2e-16

Female LFPR was forecasted from the 3rd quarter of 2020 to the 4th quarter of 2022 by using ARIMA (1,1,1) without a drift model. Forecasted values and confidence intervals at 95% levels of confidence represents in table 7.

Table 9: Forecasted values for female LFPR

Year	Point Forecast	Lo (95%)	Hi(95%)
2020 Q3	32.25183	30.01873	34.48494
2020 Q4	32.74819	29.87436	35.24155
2021 Q1	33.12615	29.89093	36.36138
2021 Q2	33.41397	29.94930	36.87864
2021 Q3	33.63313	30.01212	37.25414
2021 Q4	33.80002	30.06645	37.53359
2022 Q1	33.92710	30.10855	37.74566
2022 Q2	34.02388	30.13833	37.90942
2022 Q3	34.09756	30.15713	38.03800
2022 Q4	34.15368	30.16671	38.14064

MAPE and RMSE were calculated using test data and corresponding fitted values from the 3rd quarter of 2020 to the 2nd quarter of 2021. The corresponding RMSE was 1.37 and the MAPE value was 3.18%. Actual female LFPR in the last quarter in the training period is 31.6%. An increment in female LFPR can be expected in the next 10 quarters starting from the 3rd quarter of 2020. According to the forecast, it can be predicted 34.2% female unemployment rate by the end of 2022.

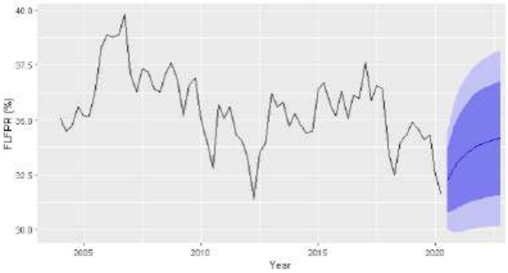

Figure 7: Time series plot of female LFPR with forecasted values

The adequacy of the predicted model was tested using residuals.

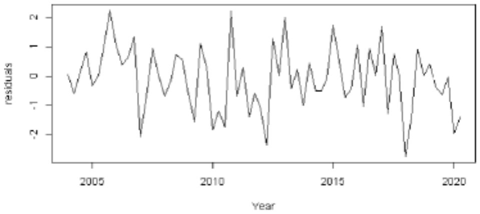

Figure 8: Plot of residuals

Figure 8 shows a plot of residuals. There is no specific pattern and residuals are randomly dispersed around the horizontal axis. All residuals are lying between +2 and -2. Box-Ljung test results were used to test the independence of residuals. Table 8 shows the results of the Box-Ljung test.

Table 10: Box-Ljung test results for residuals

Box-Ljung test	
X-squared	0.0087933
Df	1
P-value	0.9253

Since P-value (0.9253) is greater than 0.05, the null hypothesis is not rejected at a 5% level of significance by concluding that residuals are independently distributed.

Results of the Shapiro-Wilk normality test represents in table 9. since P value (0.9374) >0.05, residuals are normally distributed at a 5% level of significance.

Table 11: Results of the Shapiro-Wilk normality test

Shapiro-Wilk normality test	
W	0.9916
p-value	0.9374

4. Discussion and Conclusion

The main limitation of this study was the limited number of observations. There were only 70 observations available. Out of these 70 observations, 8 were missing. As a percentage, it was more than 10%. In time series analysis, less number of data points leads to reduce the accuracy of the forecast because it is unable to capture characteristics or past behavior of data using fewer data points. Therefore, it was decided to impute missing values without ignoring these missing values. Different imputation methods were compared other than using traditional imputation methods like mean imputation. Imputation methods were compared by error calculation (MAPE and RMSE) in between imputed value and actual value. Therefore, part of the series without missing values was selected. The exponential weighted moving average method was the best imputation method for female LFPR. Female LFPR does not show a rapid decreasing pattern. But there is a light decreasing pattern with fluctuations. It implies that the contribution to the labor force by females was reduced during the period of study. Female LFPR varies between 39.8 % (maximum value) to 30.9% (minimum value). This minimum female LFPR was obtained in 2nd quarter of 2021. The female LFPR value is predicted as 34.2% by ARIMA (1,1,1) without a drift model. It can be expected a small increment in the unemployment rate in the 2nd quarter of 2021.

References

Gunathilaka, R., 2013. *Women's participation in Sri Lanka's labor force: trends, drivers and constraints.* Colombo: ILO

Kularathna, H., 2007. Structural Change and the State of the Labour Market in Sri Lanka. *University of Colombo electronic repository,* pp. 179-220

Lechman, E. & Kaur, H., 2015. Economic Growth and Female Labor Force Participation Verifying the U-Feminization Hypothesis. *SRN Electronic,* 8(1), pp. 246-257.

Moritz, S. & Beielstein, T. B., 2017. imputeTS: Time Series Missing Value Imputation in R. *The R Journal,* 9(1), pp. 207-218.

Smarakoon, S. & Mayadunne, G., 2018. An exploratory study on low labor force participation of women in Sri Lanka. *Sri Lanka Journal of Social Sciences,* 41(2), p. 137.

Abbreviations and specific symbols

Labor force participation rate (LFPR), Akaike's information criterion(AIC), Augmented dickey fuller(ADF), Autoregressive integrated moving average(ARIMA)

Author Biographies

Ms.NN Ransinghe is a graduate of the University of Sri Jayewardenepura and pursuing MSc in Applied Statistics at the University of Colombo.

Dr. RAB Abeygunawardana (Ph.D.), Senior Lecture grade 1in the University of Colombo.