Forecasting Number of Female Births

Table of Content

Summary

For the complete Time Series Analysis and Forecasting, all the steps were followed from data selection, exploring and visualizing the series and evaluating predictability, pre-processing of data, partitioning of time series, followed by generating numerous forecasting model, comparing the results of these models and then implementing the best model(s) for forecasting of data for 8 quarters in future, and deriving conclusions.

The methods used for forecasting / model generation were: (a) Seasonal Naive (as the base for comparison of accuracy results), (b) Moving Average - Trailing (with 4 different window widths), (c) Advanced Exponential Smoothing using Holt-Winter's method, (d) Regression with Linear Trend, (e) Regression with Quadratic Trend, (f) Regression with Seasonality, (g) Regression with Linear Trend and Seasonality, (h) Regression with Quadratic Trend and Seasonality and (h) Auto ARIMA. The selection of these models was made based on the merit / demerit of each.

The forecasting results were quite promising with MAPE (Mean Absolute Percentage Error) being **2.789** and **2.809** and RMSE (Root Mean Square Error) as **115.073** and **122.954** for **Regression with Linear Trend and Seasonality** and **Auto-ARIMA: (0,0,0)(1,1,0)[4]** respectively. The data had seasonality in it with linear trend.

Introduction

Quarterly data for female births of a region in New Zealand was selected. The data was taken from the website: https://timeseries.weebly.com/data-sets.html. The reason for selection of this dataset was that the prediction of number of births is extremely important for any government and most of its constituting bodies for medium and long term planning for infrastructural requirements like schools, teachers, utilities, commodities, etc. In addition to the government, many organizations depend heavily on birth estimates especially companies like Babies-R-Us, Toys-R-Us, Diaper manufacturing companies like Pampers, Huggies, etc., Strollers and Crib manufacturing companies like Graco, Cosco, and many other companies like Gerber, Carter, etc. We had data for both males births and female births, but we selected data on female births as the whole world is talking about female empowerment, female education, female liberty, etc. We checked the predictability of data for female births and found that it has some relevant components and is predictable.

Female births' data was picked up for analysis and forecasting. The dataset had quarterly data for 10 years from 2000 - Q1 to 2012 - Q4 i.e. 48 historical data points, which was found satisfactory to prepare model and forecast the numbers of female births in the region for the next 2 years on a quarterly basis. It was decided that the accuracy for seasonal naive, a few regression models, some moving average model, some smoothing model and some ARIMA models will be checked subsequent to the development of the respective models, and then forecast for future will be done using the best model or best 2 models if 2 models are very close as far as accuracy is concerned.

Forecasting Process

Step 1: Define Goal

The goal of this analysis was to forecast the number of births in a region of New Zealand for 8 quarters in future using one of the most optimum forecasting models applying the learnings of the Time Series Analysis and Forecasting Course and utilizing R as the software tool.

Step 2: Get Data

Quarterly data for female births of a region in New Zealand was taken from the website: https://timeseries.weebly.com/data-sets.html. The data had other columns too, but the relevant column only was copied on to the csv file (Births.csv) used as input for this analysis and forecasting. After putting the data in time series dataset, the following were the head and tail rows of the time series:

```
> head(births.df)
  Quarter FemaleBirths
1  2000Q1         3070
2  2000Q2         3178
3  2000Q3         3511
4  2000Q4         3151
5  2001Q1         3070
6  2001Q2         3392
> tail(births.df)
   Quarter FemaleBirths
47  2011Q3         4247
48  2011Q4         3705
49  2012Q1         3401
50  2012Q2         3611
51  2012Q3         4287
52  2012Q4         3744
```

Step 3: Explore and Visualize Series

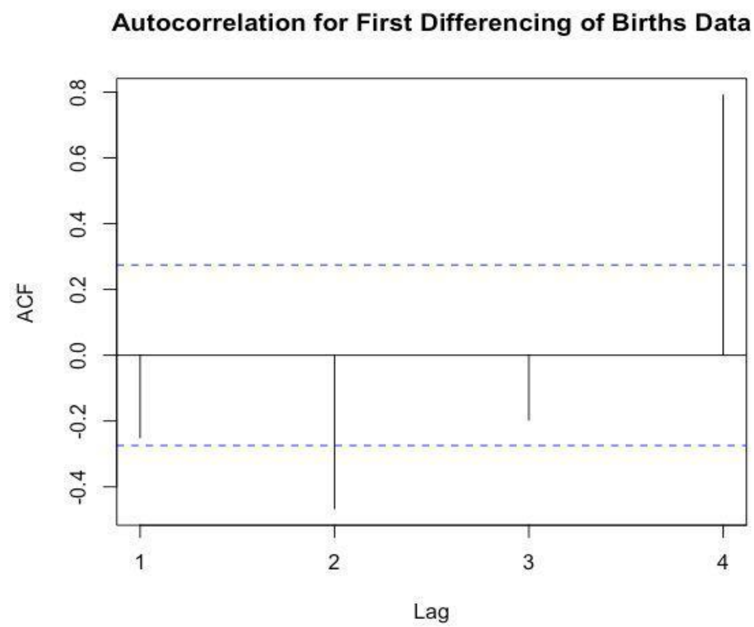
The first and foremost step in exploring and visualizing series was to see if the data is at all predictable, else the whole effort of forecasting will be a waste if the data is found to be a random walk / unpredictable. Hence, both the methods to analyze the predictability of data were applied as follows:

Evaluating Predictability : Using AR(1) model

```
> births.ar1<- Arima(births_f.ts, order = c(1,0,0))
> summary(births.ar1)
Series: births_f.ts
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1       mean
      0.0818  3556.2772
s.e.  0.1403    49.0319

sigma^2 estimated as 109956:  log likelihood=-374.57
AIC=755.14   AICc=755.64   BIC=761

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE     MASE       ACF1
Training set 0.7959748 325.1566 260.9863 -0.7754745 7.229046 1.772152 0.03155596
```

Here, the coefficient of the ar1 variable, 0.0818 which is well below 1 and even with confidence of 95%, the upper value of this coefficient will still be far below 1. Hence, it was inferred that births_f.ts is not a random walk and is predictable.

Evaluating Predictability : Using Acf() function with 1st differencing:



Using the first differencing of the historical data and Acf() function the above plot was obtained, where autocorrelation coefficients for lag 2 and lag 4 are statistically significant. Hence, it was inferred that the time series is not a random walk, rather predictable.

Next summary of data was run to see if any outliers exist due to data entry errors or otherwise:

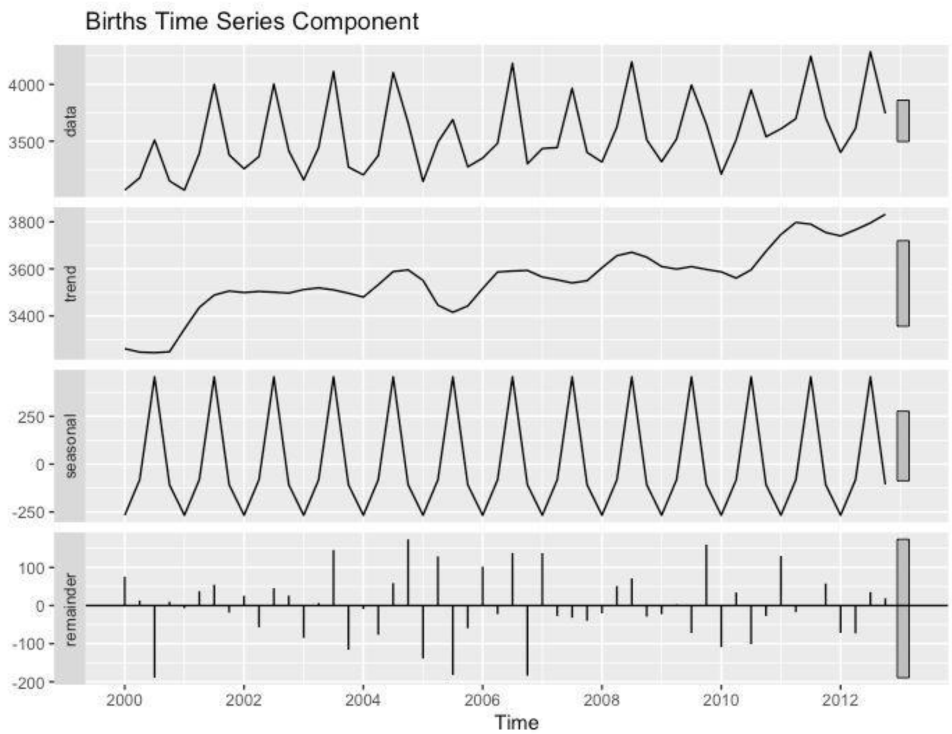
```
> summary(births.df)
    Quarter     FemaleBirths
 2000Q1 : 1   Min.   :3070
 2000Q2 : 1   1st Qu.:3319
 2000Q3 : 1   Median :3488
 2000Q4 : 1   Mean   :3557
 2001Q1 : 1   3rd Qu.:3700
 2001Q2 : 1   Max.   :4287
 (Other):46
```

It was found that the mean, median, quartiles values are all looking good and all were integers.

No abnormality was found with the data including outliers.

Time Series Components:

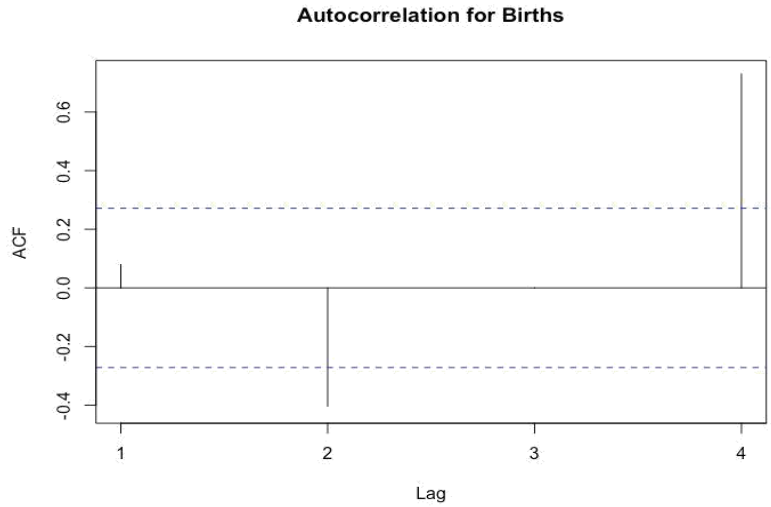
Then using stl() function various time series components were analyzed as below:



The plot above shows that there seems to be an upward trend and very strong seasonality. The data does not seem to be too levelled.

Autocorrelation of entire time series

As seen in the graph below, autocorrelation coefficients for lag 2 and lag 4 are statistically significant. Especially, autocorrelation coefficient for lag 4 shows strong seasonality, however, that for lag 3 is almost negligible.



Step 4: Preprocess Data

No data preprocessing was required as the data was found to be clean and straight-forward.

Step 5: Partitioning of Time Series

It was decided to keep 3 years data i.e. total 12 quarterly periods in validation dataset and balance 10 years' data in the training dataset, making validation dataset as around 23% of the entire time series data. The data for both the partitions are shown below:

```
> train.ts
     Qtr1 Qtr2 Qtr3 Qtr4
2000 3070 3178 3511 3151
2001 3070 3392 4000 3380
2002 3258 3365 4003 3416
2003 3160 3444 4113 3273
2004 3204 3374 4104 3662
2005 3145 3493 3690 3275
2006 3352 3482 4185 3302
2007 3436 3444 3965 3402
2008 3317 3625 4198 3513
2009 3320 3520 3994 3650
> valid.ts
     Qtr1 Qtr2 Qtr3 Qtr4
2010 3211 3513 3951 3540
2011 3609 3698 4247 3705
2012 3401 3611 4287 3744
```

Step 6: Applying Forecasting Methods

1. Seasonal Naive

```
> round(accuracy(births.snaive.pred$mean, valid.ts), 3)
            ME    RMSE     MAE   MPE  MAPE  ACF1 Theil's U
Test set 88.75 162.697 133.583 2.227 3.525 0.209     0.412
```

Seasonal Naive forecast was mainly used as a base for comparison. A table was prepared in the end comparing accuracy measures (MAPE and RMSE) for various forecasting methods. We can see baseline MAPE is 3.525 and RMSE is 162.697.

2. Moving Average - Trailing

```
> round(accuracy(ma.trailing_3, births_f.ts), 3) #best
            ME    RMSE    MAE    MPE  MAPE   ACF1 Theil's U
Test set 15.66 295.828 250.66 -0.169 6.912 -0.057     0.692
> round(accuracy(ma.trailing_4, births_f.ts), 3)
             ME    RMSE     MAE   MPE MAPE   ACF1 Theil's U
Test set 17.643 297.964 234.214 -0.16 6.37 -0.086     0.707
> round(accuracy(ma.trailing_5, births_f.ts), 3)
             ME    RMSE     MAE   MPE  MAPE   ACF1 Theil's U
Test set 24.571 255.223 198.496 0.132 5.385 -0.068     0.598
> round(accuracy(ma.trailing_6, births_f.ts), 3)
             ME    RMSE     MAE   MPE  MAPE   ACF1 Theil's U
Test set 32.961 273.128 212.323 0.344 5.728 -0.095     0.633
```

Models for Trailing Moving Averages were generated using rollmean() function with window width of 3, 4 , 5 and 6. It was sort of necessary to take window width of 4 as it was quarterly data. Accuracy measures were found using accuracy() function for all these models as shown above. As we can see the MAPE and RMSE for Training MA for window width 5, 5.385 and 255.223 respectively are the minimum amongst all Trailing MA models. The table at the end of this section compares the accuracy measures (MAPE and RMSE) for various forecasting methods including these.

3. Advanced Exponential Smoothing (Holt-Winter's Model)

Next Holt-Winter's method was used with ets() function and model = "ZZZ" to get the system selected optimum model for error trend and seasonality. The model, as shown below resulted in multiplicative error, additive trend and additive seasonality (M,A,A), but with all, alpha (exponential smoothing constant), beta (smoothing constant for trend) and gamma (smoothing constant for seasonality) as negligible (all 0.0001). It means all the 3 components, error, trend and seasonal tend to be globally adjusted and does not change over time. The MAPE is 3.953 and RMSE is 168.451 .

```
> hw.ZZZ
ETS(M,Ad,A)

Call:
 ets(y = train.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 1e-04
    beta  = 1e-04
    gamma = 1e-04
    phi   = 0.9139

  Initial states:
    l = 3229.6903
    b = 36.1011
    s = -132.7699 481.3849 -73.0672 -275.5477

  sigma:  0.0369

     AIC     AICc      BIC
546.1620 553.7482 563.0508
> round(accuracy(hw.ZZZ.pred$mean, valid.ts), 3)
              ME    RMSE     MAE   MPE  MAPE  ACF1 Theil's U
Test set 103.317 168.451 148.155 2.696 3.953 0.415     0.407
```

4. Regression Based Models

i. Regression model with linear trend

```
> summary(train.trend)

Call:
tslm(formula = train.ts ~ trend)

Residuals:
    Min      1Q  Median      3Q     Max 
-370.36 -211.75  -96.27  155.68  651.12 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3328.204     99.643  33.401   <2e-16 ***
trend          8.912      4.235   2.104    0.042 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 309.2 on 38 degrees of freedom
Multiple R-squared:  0.1044,	Adjusted R-squared:  0.08079 
F-statistic: 4.428 on 1 and 38 DF,  p-value: 0.04203
```

In the above model, though the trend coefficient is statistically significant and the p-value for F-statistic is also below 0.05, however, the R-squared values is only 0.1044. Hence, this model can not produce great forecast.

ii. Regression model with quadratic trend

```
> summary(train.quad)

Call:
tslm(formula = train.ts ~ trend + I(trend^2))

Residuals:
   Min     1Q Median     3Q    Max 
-400.1 -214.7 -104.6  126.5  628.1 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3264.0824   155.7832  20.953   <2e-16 ***
trend         18.0722    17.5235   1.031    0.309    
I(trend^2)    -0.2234     0.4145  -0.539    0.593    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 312.1 on 37 degrees of freedom
Multiple R-squared:  0.1113,	Adjusted R-squared:  0.0633 
F-statistic: 2.318 on 2 and 37 DF,  p-value: 0.1126
```

In the above model, both the trend and trend^2 coefficients are statistically insignificant, p-value for F-statistic is also 0.11 (statistically insignificant), also, R-squared is only 0.1113. Hence, this model can not be used for forecasting.

iii. Regression model with seasonality

```
> summary(train.seas)

Call:
tslm(formula = train.ts ~ season)

Residuals:
    Min      1Q  Median      3Q     Max 
-465.30  -76.95   12.95   93.87  259.60 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3233.20      51.07  63.310  < 2e-16 ***
season2       198.50      72.22   2.748   0.0093 ** 
season3       743.10      72.22  10.289 2.89e-12 ***
season4       169.20      72.22   2.343   0.0248 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 161.5 on 36 degrees of freedom
Multiple R-squared:  0.7685,	Adjusted R-squared:  0.7493 
F-statistic: 39.85 on 3 and 36 DF,  p-value: 1.569e-11
```

In this model, all the coefficients, for the 3 seasons and for F-statistic are statistically significant and R-squared is 0.7685, which is decent. Hence, this model can be used for forecasting and its equation can be given as: y_t= 3223.2 + 198.5D2 + 743.1D3 + 169.2D3, where $y_t$is the forecast for period t and D2, D3 and D4 are binary variables for Q2, Q3 and Q4.

iv. Regression model with linear trend and seasonality

```
> summary(train.trend.seas)

Call:
tslm(formula = train.ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-321.30  -80.65   -7.30   96.10  275.60 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3081.200     53.769  57.304  < 2e-16 ***
trend          8.000      1.803   4.436 8.69e-05 ***
season2      190.500     58.630   3.249  0.00256 ** 
season3      727.100     58.713  12.384 2.39e-14 ***
season4      145.200     58.851   2.467  0.01865 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 131 on 35 degrees of freedom
Multiple R-squared:  0.8518,	Adjusted R-squared:  0.8349 
F-statistic: 50.31 on 4 and 35 DF,  p-value: 4.887e-14
```

In the above model, all the coefficients, for trend, the 3 seasons and for F-statistic are statistically significant and R-squared is 0.8518, which is pretty good. Hence, this model can be used for forecasting and its equation can be given as: y_t= 3081.2 + 8t + 190.5D2 + 727.1D3 + 145.2D3,

where y_t is the forecast for period t and D2, D3 and D4 are binary variables for Q2, Q3 and Q4.

v. Regression model with quadratic trend and seasonality

```
> summary(train.quad.seas)

Call:
tslm(formula = train.ts ~ trend + I(trend^2) + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-328.74  -77.00   -5.21   95.13  247.48 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3021.1930    72.7752  41.514  < 2e-16 ***
trend         16.6024     7.3096   2.271  0.02958 *  
I(trend^2)    -0.2098     0.1728  -1.214  0.23316    
season2      190.0804    58.2382   3.264  0.00251 ** 
season3      726.6804    58.3207  12.460 3.15e-14 ***
season4      145.2000    58.4571   2.484  0.01808 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 130.2 on 34 degrees of freedom
Multiple R-squared:  0.858,	Adjusted R-squared:  0.8371 
F-statistic: 41.09 on 5 and 34 DF,  p-value: 1.833e-13
```

In the above model, the coefficients for trend, the 3 seasons and for F-statistic are statistically significant and R-squared is 0.858, which is pretty good. However, the coefficient for trend^2 is statistically insignificant with high p-value, which indicates that quadratic trend is missing in the data. But, this model can still be used for forecasting (subject to acceptable accuracy measures) and its equation can be given as: $y_t = 3021.19 + 16.6t - 0.21\ t^2 + 190.08D2 + 726.68D3 + 145.2D3$, where $y_t$is the forecast for period t and D2, D3 and D4 are binary variables for Q2, Q3 and Q4.

Comparison

The accuracy measures for the 5 regression models tried above are given below and we can see that the regression model with linear trend and seasonality has the minimum MAPE value of 2.789 and RMSE value of 115.073.

```
> round(accuracy(train.trend.pred$mean, valid.ts), 3)
              ME    RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set -32.862 294.792 240.161 -1.516 6.449 -0.04     0.672
> round(accuracy(train.quad.pred$mean, valid.ts), 3)
             ME    RMSE    MAE  MPE  MAPE ACF1 Theil's U
Test set 91.061 321.291 232.61 1.81 6.054 0.03     0.766
> round(accuracy(train.seas.pred$mean, valid.ts), 3)
             ME    RMSE     MAE   MPE  MAPE  ACF1 Theil's U
Test set 198.85 237.556 206.767 5.275 5.497 0.447     0.591
> round(accuracy(train.trend.seas.pred$mean, valid.ts), 3)
            ME    RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set -9.15 115.073 101.767 -0.353 2.789 0.372      0.26
> round(accuracy(train.quad.seas.pred$mean, valid.ts), 3)
              ME    RMSE     MAE  MPE  MAPE  ACF1 Theil's U
Test set 107.227 175.444 154.974 2.77 4.126 0.464     0.422
```

5. Auto-Arima

Last but not the least, Auto-ARIMA method was used for model development and the results were sort of surprising. The model summary is presented below. The result though is a seasonal ARIMA model with 3 parentheses, with AR component with no trend, no differencing and no seasonality (p,d,q = 0) and with AR seasonal component of order 1 (P=1), with 1 differencing (D=1) and no MA component (Q=0). Last parenthesis shows quarterly data (4). There is a drift.

```
> summary(train.auto.arima)
Series: train.ts
ARIMA(0,0,0)(1,1,0)[4] with drift

Coefficients:
         sar1   drift
      -0.6022  9.7538
s.e.   0.1367  4.3095

sigma^2 estimated as 26601:  log likelihood=-234.35
AIC=474.7   AICc=475.45   BIC=479.45

Training set error measures:
                   ME    RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 8.090206 150.368 110.4176 0.1203728 3.113595 0.7385793 0.0872716
```

```
> round(accuracy(train.auto.arima.pred, valid.ts), 3)
                  ME    RMSE     MAE    MPE  MAPE  MASE  ACF1 Theil's U
Training set   8.090 150.368 110.418  0.120 3.114 0.739 0.087        NA
Test set     -24.487 122.954 103.071 -0.771 2.809 0.689 0.415     0.301
```

Equation for the above model can be given as: $y_t = 9.738 - 0.6022\ (y_{t-1} - y_{t-5})$.

Also, we can see that MAPE is 2.809 and RMSE is 122.954 for validation data, which is pretty low compared to the other models.

Step 7: Evaluation and Performance Comparison

In the next step of forecasting process, accuracy measures for various forecasting methods tried above were compared. For moving average method, the best of the 4 models i.e. k=5 window width was taken for comparison. And for regression models, the best of the 5 models i.e. regression with linear trend and seasonality was taken for comparison:

```
> round(accuracy(births.snaive.pred$mean, valid.ts), 3) #1: Seasonal Naive
            ME    RMSE     MAE   MPE  MAPE  ACF1 Theil's U
Test set 88.75 162.697 133.583 2.227 3.525 0.209     0.412
> round(accuracy(hw.ZZZ.pred$mean, valid.ts), 3) #2 : Advanced Exponential Smoothing (Holt-Winter's Model)
             ME    RMSE     MAE   MPE  MAPE  ACF1 Theil's U
Test set 103.317 168.451 148.155 2.696 3.953 0.415     0.407
> round(accuracy(ma.trailing_5, births_f.ts), 3) #3 : Moving Averages
             ME    RMSE     MAE   MPE  MAPE   ACF1 Theil's U
Test set 24.571 255.223 198.496 0.132 5.385 -0.068     0.598
> round(accuracy(train.trend.seas.pred$mean, valid.ts), 3) #4 : Regression model with linear trend and seasonality
            ME    RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set -9.15 115.073 101.767 -0.353 2.789 0.372      0.26
> round(accuracy(train.auto.arima.pred$mean, valid.ts), 3) #5 : Auto Arima
              ME    RMSE     MAE    MPE  MAPE  ACF1 Theil's U
Test set -24.487 122.954 103.071 -0.771 2.809 0.415     0.301
```

Model	**MAPE**	**RMSE**	**Rank for forecasting**
Seasonal Naive	3.525	162.697	III
Holt-Winter’s Advanced Exponential Smoothing (M,A,A)	3.953	168.451	IV
Moving Averages - Trailing (k=5)	5.385	255.223	V
Regression with Linear Trend and Seasonality	2.789	115.073	I
Auto-ARIMA: (0,0,0)(1,1,0)[4]	2.809	122.954	II

From the above comparison table, it is evident that the best 2 models for forecasting are ‘Regression with Linear Trend and Seasonality’ and ‘Auto-ARIMA’. The difference in MAPE of the above 2 models is 0.02 only.

Step 8: Implementing Forecasts

As discussed above, it was decided to forecast for the 2 years in the future using both the models which were near best. The forecast, taken for the entire dataset (training and validation periods combined) as produced are given below:

1. Regression model with linear trend and seasonality on entire dataset

```
> round(all.trend.seas.pred$mean, 0)
     Qtr1 Qtr2 Qtr3 Qtr4
2013 3498 3696 4243 3687
2014 3530 3728 4275 3719
```

2. Auto-Arima on entire dataset

```
> round(auto.arima.pred$mean, 0)
     Qtr1 Qtr2 Qtr3 Qtr4
2013 3517 3699 4245 3742
2014 3567 3734 4341 3808
```

Conclusion

After utilizing 12 different models for forecasting (including 4 models with different window widths for Moving Averages and 5 models for regression), it was concluded that 'Regression model with linear trend and seasonality' was the best model for forecasting with a neck to neck competition with 'Auto-ARIMA' model. However, as Auto-ARIMA model resulted in many parameters as 0, it was felt that perhaps 'Regression model with linear trend and seasonality' should be used for forecasting the number of female births in the region.

The forecasting results were quite promising with MAPE (Mean Absolute Percentage Error) being **2.789** and **2.809** and RMSE (Root Mean Square Error) as **115.073** and **122.954** for **Regression with Linear Trend and Seasonality** and **Auto-ARIMA: (0,0,0)(1,1,0)[4]** respectively. The data had seasonality in it with linear trend.

Moving Average is usually used for data that lacks trend and seasonality. Since the data used in the project had liner trend and seasonality, moving averages method was not expected to give great results. As the data lacked quadratic or polynomial trend, quadratic trend with seasonality could not be used beneficially. Seasonal Naive was used as a base for comparison only. Holt-Winter's model could have produced great results with trend and seasonality in the data, but found to have higher MAPE and RMSE as compared to better models.

This project was a great learning opportunity for the group as many challenges were faced in accomplishing the goals and overcoming those challenges enhanced the understanding on the subject. Initially data on number of natural and planted forests in a country were taken but after initial analysis it was found that the data is not predictable and hence, had to be dropped. Then the current dataset was used.

Bibliography

1. https://timeseries.weebly.com/data-sets.html: Data was downloaded data from this website
2. PPTs, R Codes and Lecture Videos of Course ‘BAN673: Time Series Analytics and Forecasting’ by Prof. Zinovy Radovilsky.