Time series analysis of births in Catalunya (1975-2006)

Julian Ibarguen

20/06/2020

Contents

Project scripts and report available here

0.1 Introduction

We fitted an ARIMA model in the series of birth in Catalunya from February, 1975 to December, 2006. The series has a monthly periodicity comprising a total of 383 observations. We first performed an exploration of the series assessing for stationary; afterwards we performed a series of transformation on the series to achieve an stationary series, which helped us to define the parameters for the model; finally, with the ifnromation of the previous steps, we fitted an ARIMA model for prediction of the next 50 periods.

Note the purpose of this work was to show the logic of a time series analysis. On real conditions a train set shoudl have been split form the original data to fit the model, and afterward us holdout for the prediction. Here we will predict the next 50, despite not having holdout to test our prediction.

0.2 Exporatory analysis of the original series

We observe the births series is non-stationary with a overall decreasing tendency and a seasonal cycle of 12 months (Figure1). In practical terms we can say the simple Autocorrelation Function (ACF) allow us to observe if the series has trend, while the partial ACF allow us to assess for the order of the series. As we observe from the correlation plots (Figure 2):

- From the simple ACF we can observe a decreasing trend corresponding with the trend observe in the series sequence. We observe the progressive decreasing trend of the coefficients, which would be pointing at a MA serie of higher order, as a MA model of order one should have cut that trend in a more relevant and earlier manner. Furthermore, we observe that every 12 lags, the decreasing trend breaks, showing a higher coefficient than the previous one. This is showing us the seasonality of the series.
- From the partial ACF we observe the first two lag have are significant and progressively decreasing showing the second order of an AR model. Form the second lag the trend breaks and turn chaotic without a clear pattern, thus pointing to an Moving Average (MA) model. Further we can observe the seasonal patter at month 12 and 24.

After all, we can observe a general hybrid behavior between an AR and a MA models with order AR(2) and MA(3), and with seasonality.

Finally we checked for heterocedasticity with Bartlet's test (Table 1), resulting in a significant p-value of 0 we reject the H_0 of homogeneity of variances assuming an heterocedastic series.

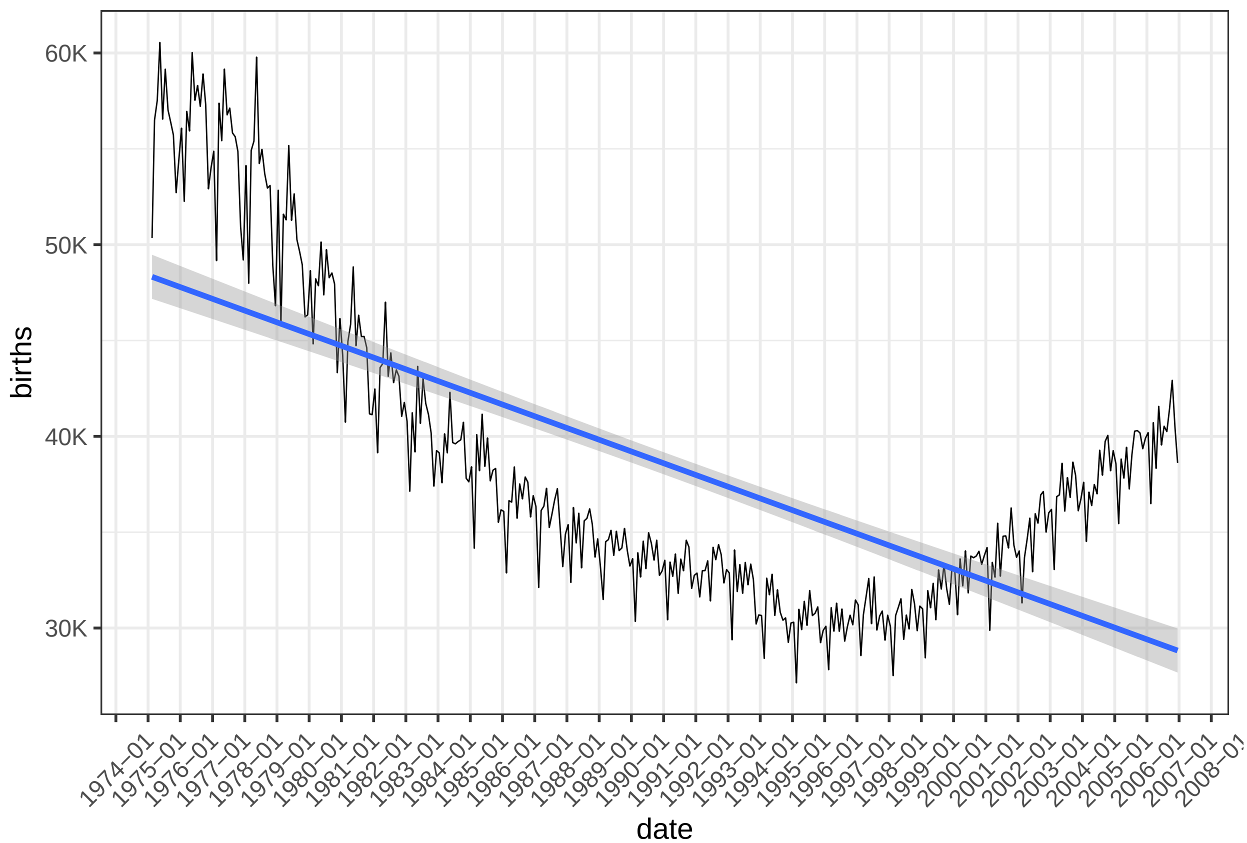
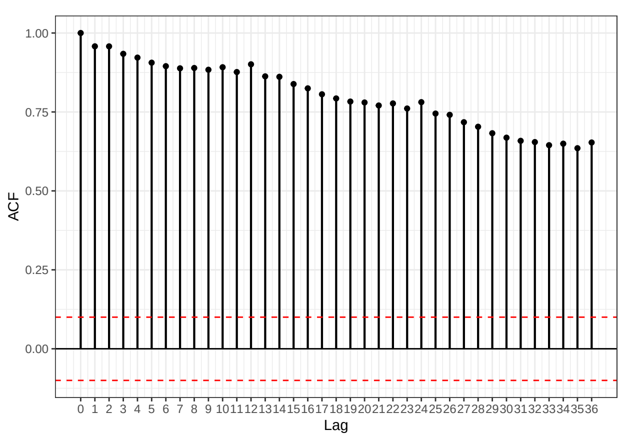
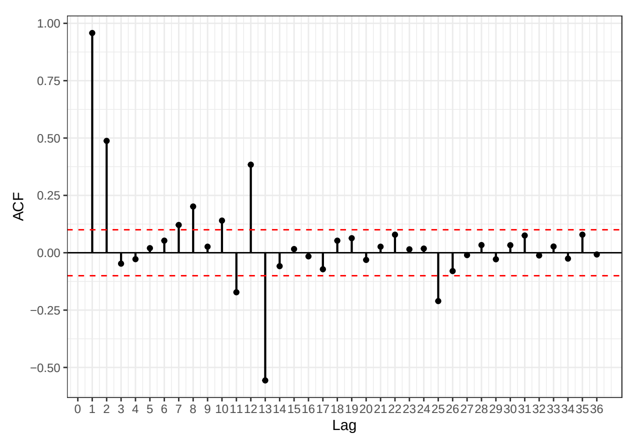


Figure 1: Birth Seires Sequence



(a) Simple autocorrelation function



(b) Partial autocorrelation function

Figure 2: Plots of the Autocorrelation function for Birth Series

Table 1: Bartlett test of homogeneity of variances of Births Series

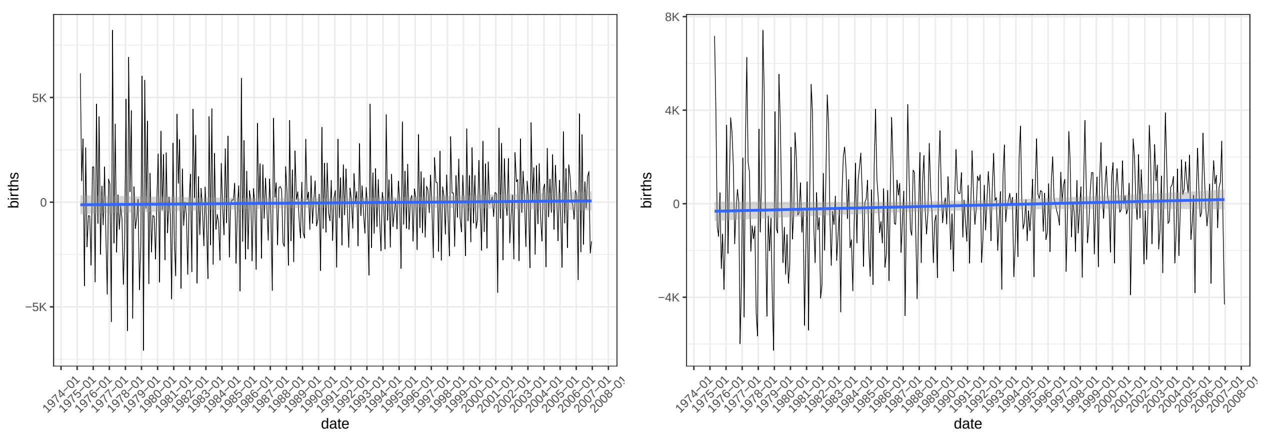
parameter	**value**
statistic.Bartlett's K-squared	91.081
parameter.df	31.000
p.value	0.000

0.3 Tranformation: towards an stationary serie

We started by removing the **non-seasonal trend** of the series with the following formula for differentiation

$$z_t = z_t - z_{t-d}$$

being d a number of periods. To proceed, we test to differentiate de series with one and two lags. We tested the correlation between a equivalent numeric value of data and the births series. We assumed the lag differentiation that provided a lower correlation coefficient would be the best differentiation. One lag differentiation resulted in a correlation coefficient of 0.026 while two lag differentiation had a coefficient of 0.078. Therefore, one lag allow us to remove the non seasonal trend and get closer to an stationary series; the same can be observe in Figure 3. A parameter $d = 1$ would seem to fit the model for non seasonal diferentiation.



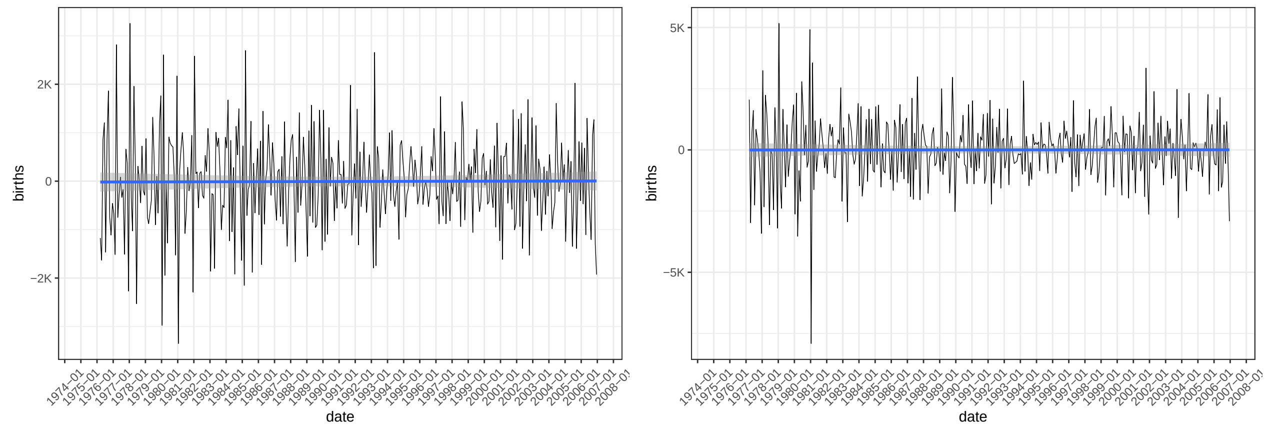
(a) One lag diferentiation (d = 1)

(b) Two lag differentiatin (d = 2)

Figure 3: Non-seasonal Differentiation of Birth Series

Once removed non-seasonal trend, we proceed to remove the **seasonal trend**. As we have seen in the exploratory the series have a seasonal cycle of 12 months. Therefore, We test seasonal differentiation with one and two seasonal cycles ($d = 12$ and $d = 24$ respectively). As we observe on Figure 4 the differentiation with the lag of one seasonal cycle (12 months) seems to offer a more stationary series than two. Furthermore, applying a seasonal differentiation of 2 cycles would lead us to lose 24 observations from the data. A parameter $d = 1$ would seem to fit the model for seasonal differentiation.

As we have seen previously in the exploratory analysis the series has some heterocedasticity. Therefore we test again for homocedasticity with Bartlett's tes to see if the current transformations have improved. We observe a p-value of 0.00055, which still is significant, thus we cannot reject heterocedasticity in our series (Table 2). However, we have seen it has nonetheless improved from the original series, moving towards a more stationary series. As logarithmic or exponential transformation would imply a large lost of observation as differentiation has returned approximately 50% of negative values, we refrain from performing further logarithmic or exponential to the series to tackle heterocedasticity further.



(a) One lag diferentiation (d = 12) (b) Two lag differentiatin (d = 24)

Figure 4: Seasonal Differentiation of Birth Series

Table 2: Bartlett test of homogeneity of variances of Differentiated Births Series

parameter	**value**
statistic.Bartlett's K-squared	61.853
parameter.df	30.000
p.value	0.001

After all, we observe the series sequence appears significantly more stationary that the original serie. This also reflect on the ACF plots that have turn less patterned and chaotic (Figures 5 and 6).

0.4 Fitting ARIMA model

We have observed that to obtain an stationary series differentiation values (d) for non-seasonal or seasonal would be 1. Further we have seen that the order for the AR part model is likely to be two, while the order of an MA is also likely to be greater order (3). Therefore from the exploratory analysis we would choose an ARIMA (2,1,3)(0,1,1). Nonetheless, we will test different models and choose the one that offers:

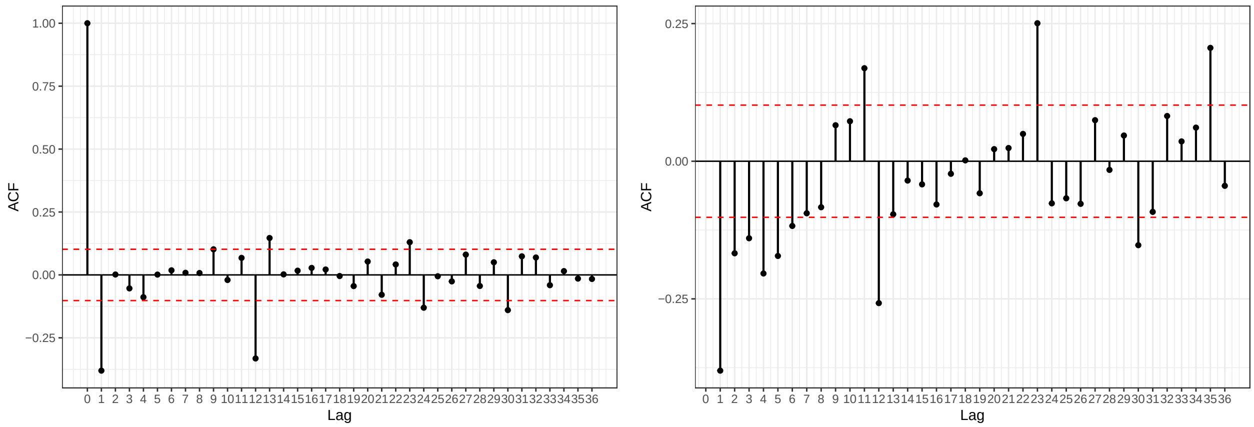
1. necessarily whose ACF residuals are not significant (white noise series)
2. All the confidence interval of 95% level does not includes 0
3. Has the lowest RMSE as priority measure
4. Has the lowest AIC

All in all we tested the following models:

Table 3: Goodness of fit measures for ARIMA models

model.desc	**sigma**	**logLik**	**AIC**	**BIC**	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**
ARIMA(2,1,3)(0,1,1)[12]	870.191	-3033.746	6081.491	6108.886	28.938	848.332	648.521	0.122	1.701	0.366	0.019
ARIMA(2,1,2)(0,1,1)[12]	895.975	-3041.360	6094.721	6118.202	27.003	874.668	671.115	0.119	1.759	0.379	0.004
ARIMA(3,1,2)(0,1,1)[12]	897.259	-3041.383	6096.766	6124.160	27.028	874.720	671.240	0.119	1.759	0.379	0.005
ARIMA(2,1,3)(0,1,2)[12]	879.270	-3033.855	6083.711	6115.019	23.585	856.005	660.484	0.105	1.746	0.373	0.051

Observing the results, we confirm that the model that we expected from our exploratory analysis ARIMA(2,1,3)(0,1,1)[12] was the one that offers the best fit among the four models we have tested (Table 3). It resulted in an RMSE of 848 and AIC of 6,081.491. Furthermore, we observe that the coefficients confidence interval does not include 0, thus all becoming significant and supporting a good fit. Finally, when we explore



(a) Simple autocorrelation function

(b) Partial autocorrelation function

Figure 5: Plots of the Autocorrelation function for Differentiated Birth Series

at the residual we have obtained a non-significant Ljung-Box test with a p-value of 0.239 (Table 5), rejecting the H_0 that the residuals are no *white noise serie*. The ACF of the residuals shows not significance for any of the lags and its distribution is approximately normal (Figure 7). All these justifies the fit of our model.

Table 4: Coeficients and confidence intervals for ARIMA models

model.desc	**parameter**	**coeficient**	**2.5 %**	**97.5 %**
ARIMA(2,1,3)(0,1,1)[12]	ar1	1.739	1.724	1.753
ARIMA(2,1,3)(0,1,1)[12]	ar2	-0.989	-1.004	-0.974
ARIMA(2,1,3)(0,1,1)[12]	ma1	-2.375	-2.440	-2.310
ARIMA(2,1,3)(0,1,1)[12]	ma2	2.070	1.952	2.188
ARIMA(2,1,3)(0,1,1)[12]	ma3	-0.604	-0.670	-0.538
ARIMA(2,1,3)(0,1,1)[12]	sma1	-0.741	-0.817	-0.665

Table 5: Ljung-Box test

method	**statistic**	**p.value**	**parameter**
Ljung-Box test	5.509	0.239	4

With the model fitted, we are ready to make the prediction of the next 50 observations. We can see how the prediction follow the trend of the series and captures also its seasonality and trend characteristics (Figure 8). Nonetheless, it cna also be observed how the accuracy of the predicction reduces the further the prediction is from the las observation of the original series. The predicted values for the next 50 observations are offered in Table 6.

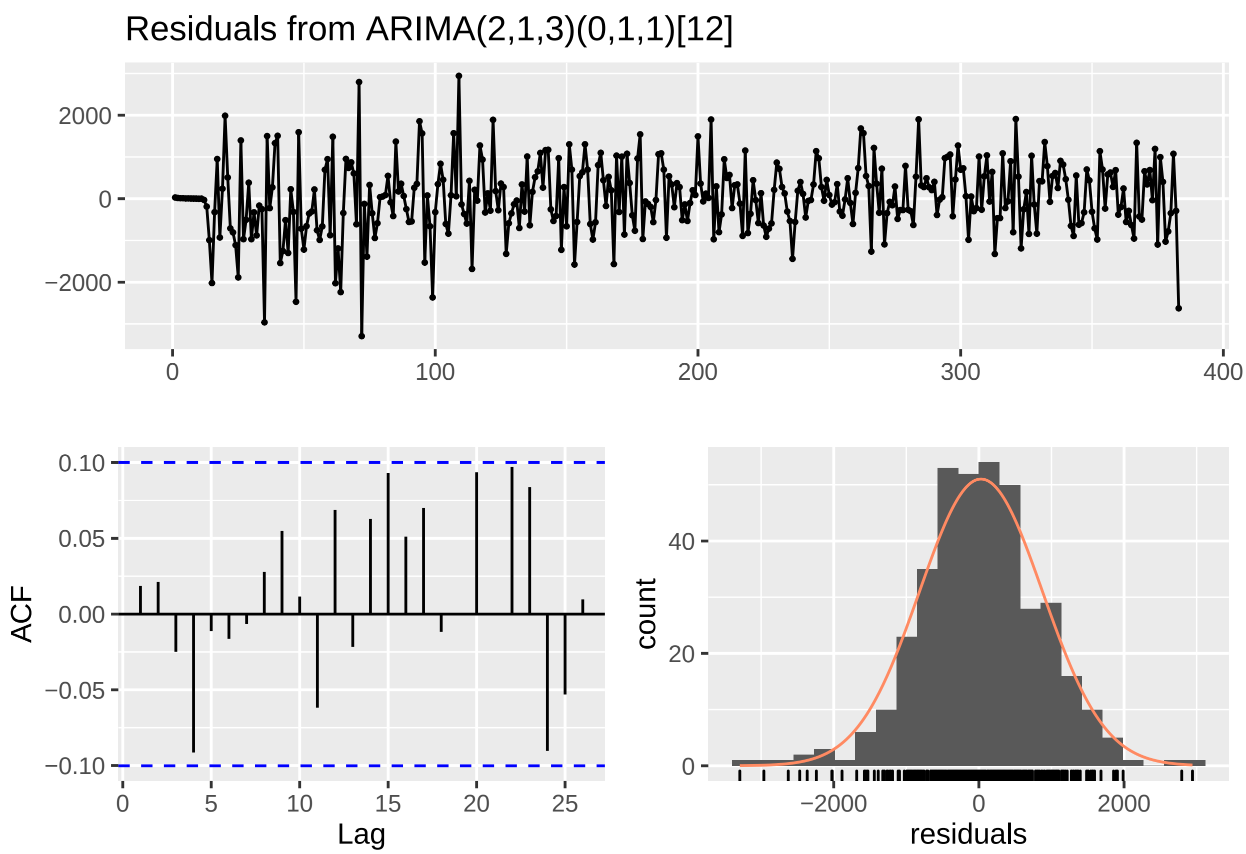


Figure 6: Birth Seires Sequence

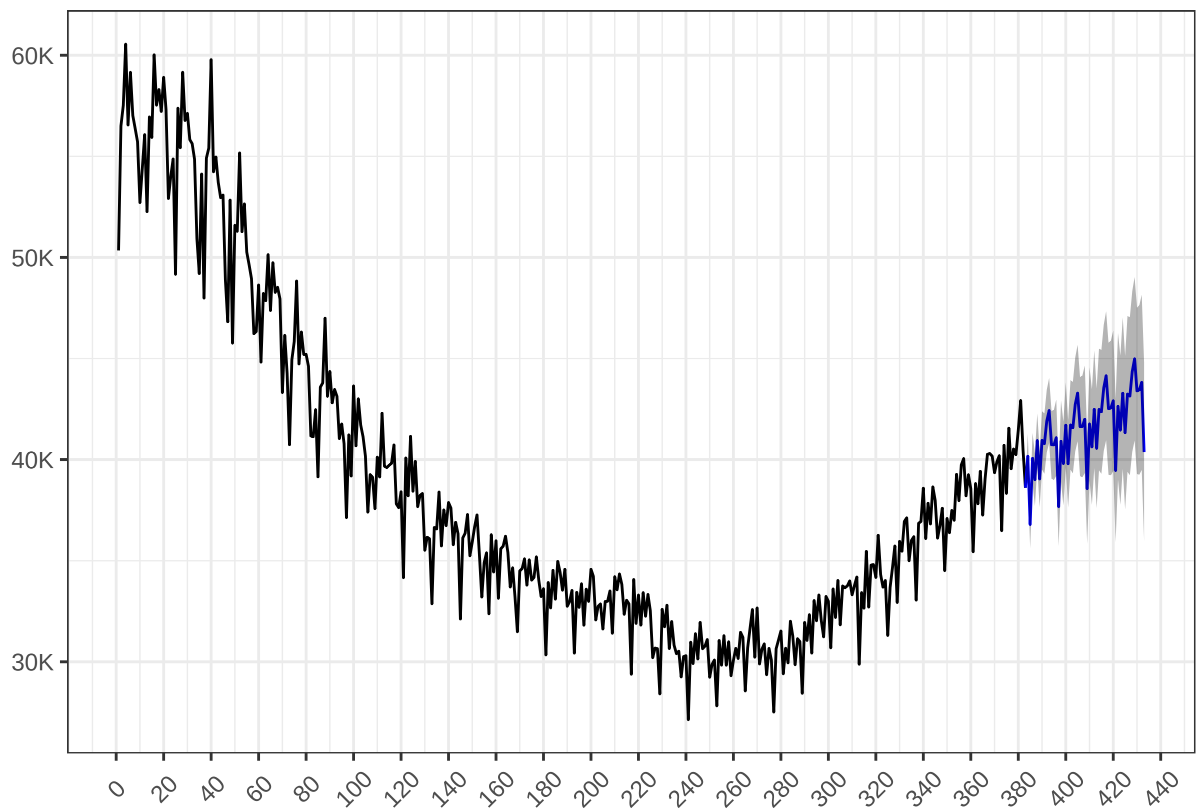


Figure 7: Birth Seires Sequence with Predicted values

Table 6: Predicted values for the next 50 observations

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
384	40173.65	39055.63	41291.68	38463.78	41883.52
385	36795.35	35603.93	37986.77	34973.23	38617.47
386	40074.25	38822.74	41325.76	38160.23	41988.27
387	39010.14	37707.86	40312.42	37018.48	41001.81
388	40942.11	39594.10	42290.13	38880.50	43003.72
389	39040.22	37647.44	40433.01	36910.15	41170.30
390	40955.71	39515.70	42395.73	38753.40	43158.03
391	40791.21	39299.14	42283.29	38509.29	43073.14
392	41907.89	40358.23	43457.56	39537.89	44277.90
393	42428.65	40817.07	44040.24	39963.95	44893.36
394	40732.61	39057.66	42407.55	38170.99	43294.22
395	40726.64	38990.49	42462.79	38071.43	43381.86
396	41086.13	39201.32	42970.94	38203.56	43968.70
397	37677.06	35719.67	39634.46	34683.49	40670.64
398	40914.76	38896.05	42933.48	37827.40	44002.13
399	39809.46	37738.21	41880.71	36641.75	42977.16
400	41710.55	39592.08	43829.01	38470.63	44950.46
401	39795.72	37631.60	41959.83	36485.99	43105.45
402	41719.26	39507.60	43930.91	38336.82	45101.69
403	41581.55	39317.77	43845.33	38119.39	45043.71
404	42736.85	40415.01	45058.70	39185.90	46287.81
405	43298.26	40912.93	45683.59	39650.21	46946.31
406	41634.68	39182.79	44086.57	37884.84	45384.52
407	41644.96	39126.99	44162.94	37794.05	45495.87
408	42000.59	39342.37	44658.80	37935.19	46065.98
409	38568.72	35833.36	41304.09	34385.34	42752.11
410	41770.62	38968.85	44572.39	37485.69	46055.55
411	40625.60	37766.50	43484.71	36252.98	44998.22
412	42493.07	39582.63	45403.52	38041.93	46944.21
413	40559.07	37599.63	43518.50	36033.00	45085.13
414	42482.51	39472.86	45492.17	37879.65	47085.38
415	42363.62	39299.53	45427.71	37677.50	47049.74
416	43551.72	40427.16	46676.29	38773.11	48330.33
417	44151.54	40960.42	47342.67	39271.13	49031.96
418	42522.32	39260.44	45784.20	37533.70	47510.93
419	42554.32	39220.90	45887.75	37456.29	47652.36
420	42913.74	39442.32	46385.17	37604.66	48222.83
421	39466.99	35912.90	43021.08	34031.47	44902.51
422	42639.24	39012.89	46265.59	37093.22	48185.26
423	41457.41	37768.19	45146.64	35815.23	47099.60
424	43290.21	39544.78	47035.64	37562.06	49018.35
425	41332.31	37533.82	45130.81	35523.01	47141.62
426	43248.54	39396.44	47100.64	37357.26	49139.82
427	43140.71	39231.24	47050.18	37161.69	49119.72
428	44355.19	40382.41	48327.98	38279.34	50431.04
429	44989.94	40947.35	49032.52	38807.34	51172.54
430	43395.34	39277.93	47512.75	37098.31	49692.37
431	43453.01	39258.91	47647.10	37038.69	49867.32
432	43822.79	39490.76	48154.81	37197.53	50448.04
433	40368.67	35948.43	44788.91	33608.50	47128.84