STA457 Final Project

David Pham

17/12/2021

Abstract

The purpose of this project is to provide an in-depth analysis of monthly births in the U.S. We use historical data from 1948-1979 of live monthly births in the U.S to create a stationary series by differencing, detrending and deseasoning. After performing various diagnostics and model selections, we obtain a SARIMA model and use it to forecast future values, as well as perform spectral analysis. We conclude that birth rates have gone down throughout the years, and the present also reflects these results.

Keywords: time series, births, seasonal ARIMA, model selection, forecasting, spectral analysis

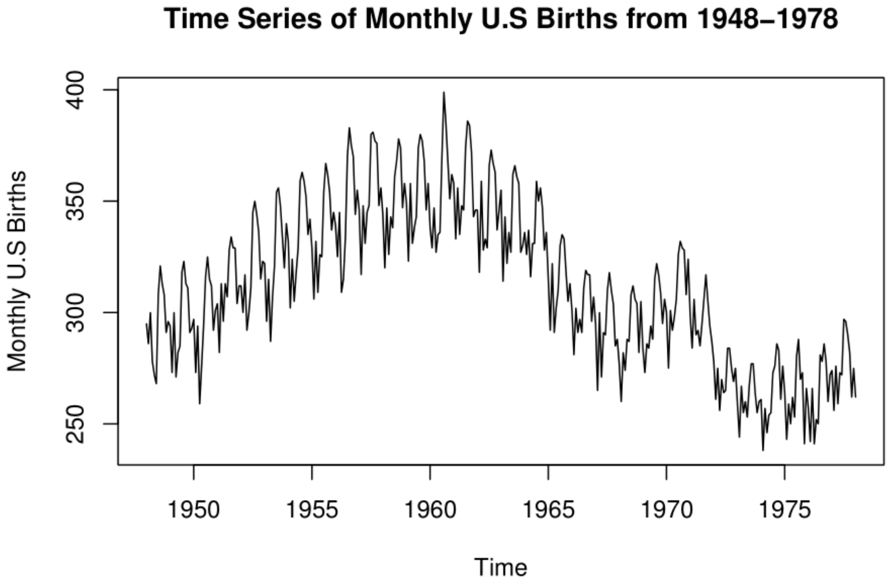
Introduction

Birth rates are a very interesting statistic to observe, as they can provide many insights about a country (and even the world), including economic status, education, social factors, and so much more. Throughout the years, birth rates have been slowly declining and are at their lowest since 1979 (BBC, 2021). The PRB (Population Reference Bureau) suggests that fertility patterns in the U.S are primarily affected by not only economic recessions such as the current pandemic, but other factors including the technological and overall development of a country (PRB, 2020). We are interested in seeing if this claim and reasoning holds by using data of the past to forecast into the future. Hence, the purpose of this project is to provide an analysis, as well as developing a suitable model to forecast the number of monthly births in the U.S. To attempt this, we use the birth dataset in the astsa package in R, which contains 373 observations of monthly live births in the U.S from January 1948 to January 1979. Throughout this report, we will be observing the data and modifying it to achieve stationarity, picking a suitable (S)ARIMA model by using necessary diagnostics and model selection, and use this final model to forecast the claims made above. In order to do the last step, we will take the last 12 observations as our testing data and perform one-step forecasts.

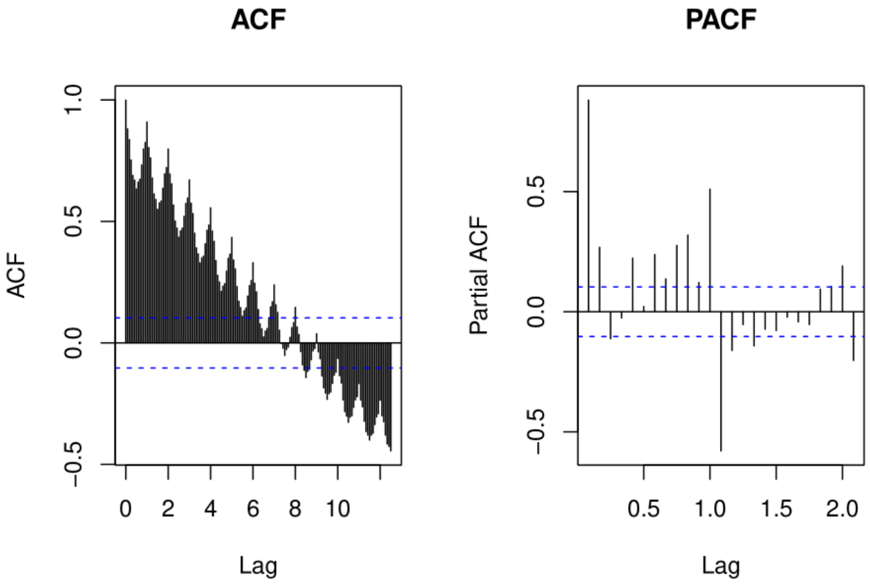
Data Analysis and Methods

Stationarity

First, let us examine the initial time series plot of monthly births from 1948-1978. The last year is omitted since we are using that data as the testing data for forecasting.



By first glance, we can observe that the data is not stationary since the mean and variance of the time series seem to be changing over time. The upward trend of the first half of the data is presumably because of the baby boom (History, 2021), while there are decreasing fluctuations in the second half. Moreover, there appears to be seasonality in the data due to roughly constant peaks throughout the series. We can also check the autocorrelation function to appropriately assert seasonality and non-stationarity:



Upon looking at the ACF and PACF of the data, we can clearly see non-stationarity due to the general decreasing trend in the ACF, as well as the points spiking higher than the blue dotted range from both graphs. We will have to transform our data to make it stationary, as well as taking seasonality and trend into account.

Next, we will employ the Box-Cox Transformation to stabilize the variance, since it varies with time (Zach, 2020). The variance of this time series is 1240.041. Although this transformation is also commonly used for normalizing data, we do not necessarily have this problem here since the data already looks normal.

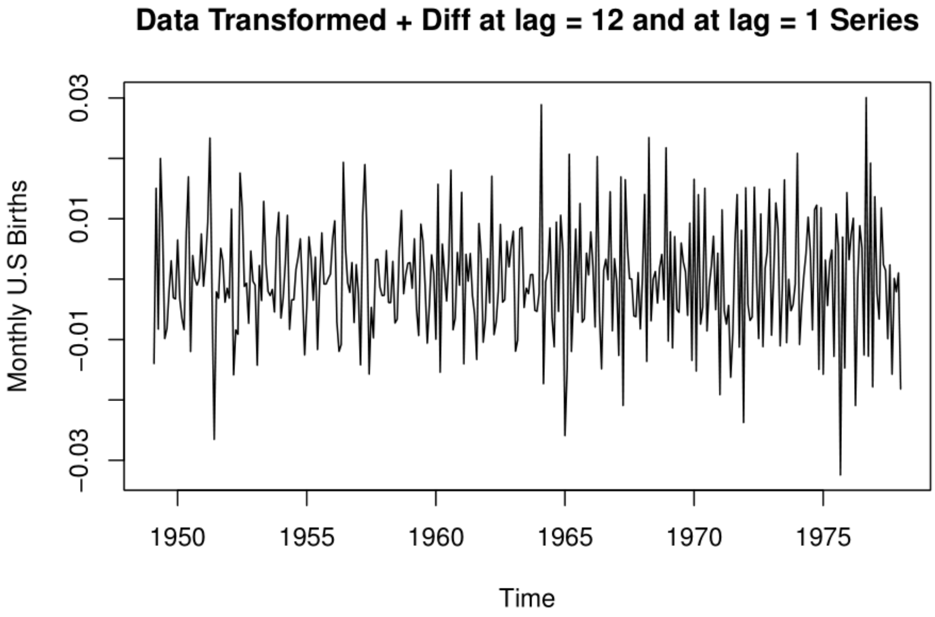
With the help of the boxcox() function, we transform our data with the equation

$$y(\lambda) = \frac{y_i^{\lambda} - 1}{\lambda}$$

where $\lambda = -0.2$ (Zach, 2020). Checking the variance of the newly transformed data, it is 0.0013!

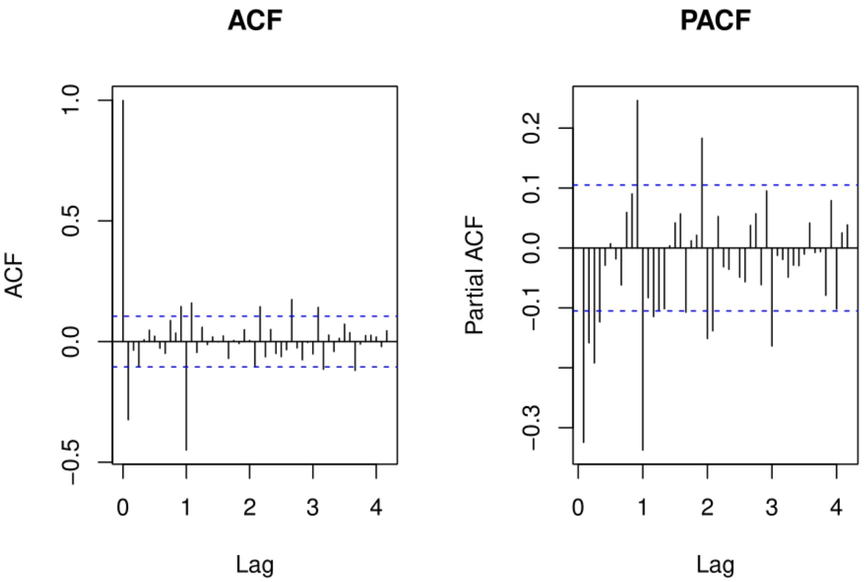
Finally, we will try to use differencing in order to stabilize the mean and make our entire series stationary. Looking at the previous and updated ACFs/PACFs, as well as decomposing the data, there seems to be seasonality at lag 1 and 12; hence, we will take the seasonal difference, and then first difference (Forecasting: Principles and Practice, 2021).

After differencing at lag = 12, we see an even smaller variance (almost 0) once we difference even further at lag = 1. We also see significant improvements in the shape of our data:



After having detrended, deseasoned, normalizing, and stabilizing the mean and variance of our data, we affirm that we have a stationary series. Additionally, we use the Augmented Dickey-Fuller (ADT) test for stationary hypothesis testing to obtain a p-value < 0.01; hence, we reject the null-hypothesis and assert that the series is stationary (Zach, 2021). Finally, we will take a look at the final ACF and PACF and determine a few model candidates.

Model Building



Observing the ACF, it appears that at the seasons, the ACF tails off at lag 1s (s = 12). Furthermore, the PACF also tails off at lags 1s, 2s, 3s, 4s; hence, these results seem to imply an SMA(1), where P = 1, Q = 1, D = 1 in the seasonal component (Shumway, 2017). We took the seasonal and first difference, D = d = 1. In the non-seasonal lags, we can see that the PACF is tailing off. Furthermore, if we ignore seasonal lags, the ACF cuts off after lag = 2, which suggests a non-seasonal MA(2) component (Shumway, 2017). Hence, we will test out a few SARIMA models, but our ideal guess is of the form (0, 1, 2)(1, 1, 1)$_{12}$.

Results

Model Selection, Diagnostics, and Interpretation

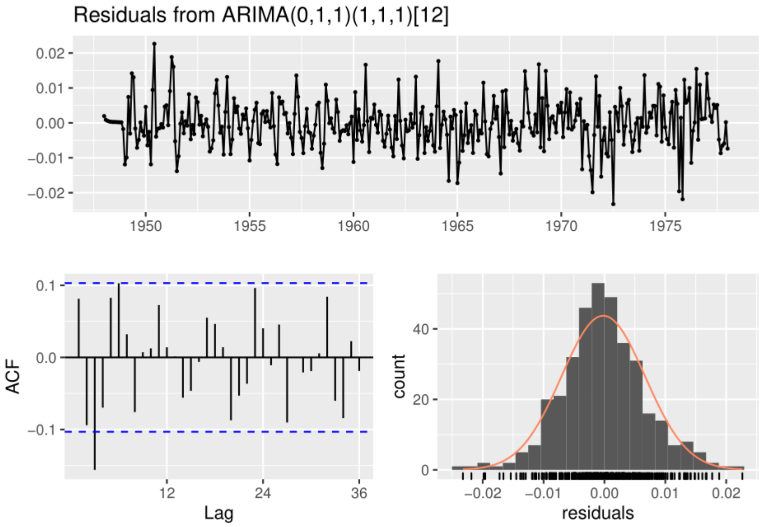
Here is a table of our 8 candidate SARIMA models:

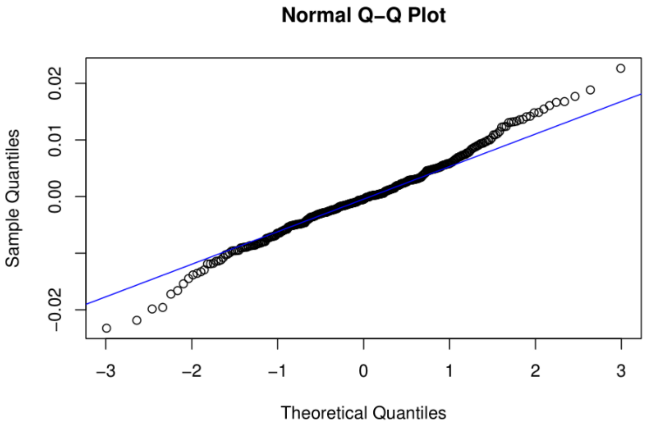
Table 1: AICc Values for SARIMA Models

Model	AICc
SARIMA(0,1,1)(0,1,0)[12]	-1786.7
SARIMA(0,1,1)(0,1,1)[12]	-2089.8
SARIMA(0,1,1)(1,1,0)[12]	-1973.5
SARIMA(0,1,1)(1,1,1)[12]	-2157.5
SARIMA(0,1,2)(0,1,0)[12]	-1841.6
SARIMA(0,1,2)(0,1,1)[12]	-2143.3
SARIMA(0,1,2)(1,1,0)[12]	-2029.2
SARIMA(0,1,2)(1,1,1)[12]	-2211.2

It actually appears that the best three models are the SARIMA $(0,1,1)(1,1,1)_{12}$ (we'll call this Model 1), $(0,1,2)(1,1,1)_{12}$ (Model 2) and $(0,1,2)(0,1,1)_{12}$ (Model 3) as they have the lowest AICc values by a tiny margin. We will use these final models in our diagnostics.

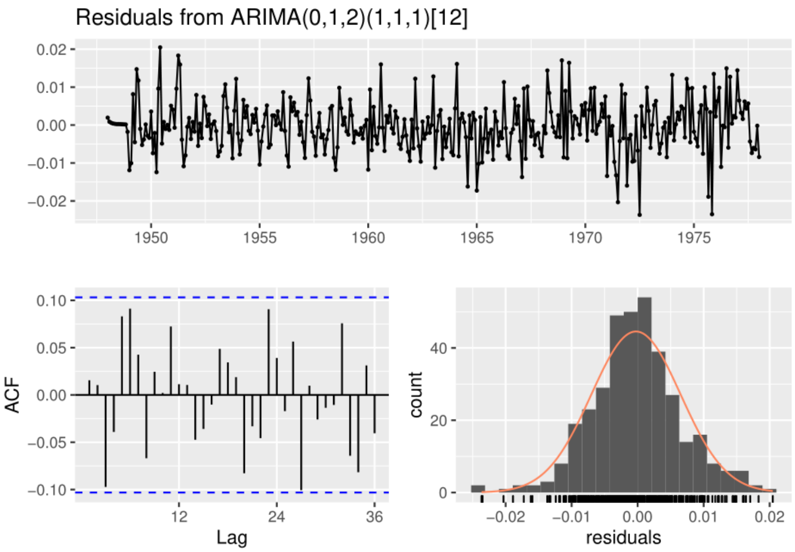
For Model 1, the mean and variance of the residuals are both almost 0. Upon plotting the residual diagnostics:

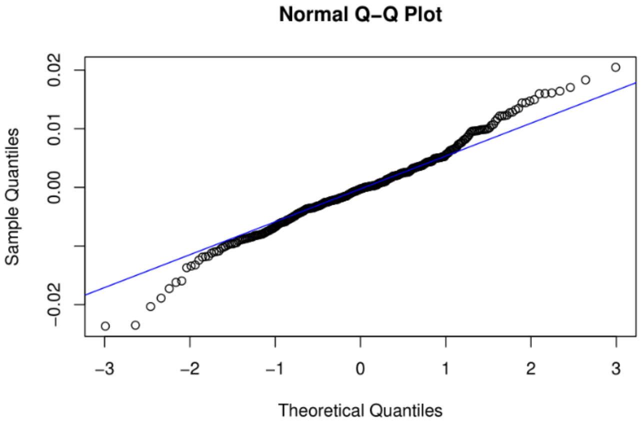




Model 1 fails the Ljung-Box test, as the p-value is 0.007; hence we believe that the time series is autocorrelated by rejecting the null hypothesis. Furthermore, some of the lags go past the blue confidence band, which is not typically ideal for white noise residuals. In the Q-Q plot, the tails seem to deviate quite a bit from the blue line. Due to these reasons, we will scrap Model 1.

Next, Models 2 & 3 also have mean and variance close to 0. They are very similar because they only differ by a seasonal regressive order, and so their diagnostics are alike:

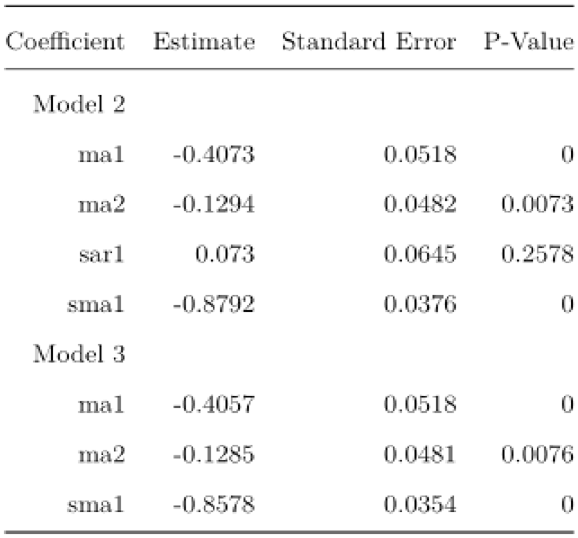




The Ljung-Box test gives us a p-value > 0.1 for both, where we fail to reject the null hypothesis; in other words, the series are not autocorrelated (which is what we want!). Furthermore, the residuals look like white noise in the ACF, and they look normally distributed. Most of the points on the Q-Q plot are also touching the blue line, with slight deviations in the tails. Overall, these two SARIMA models are good candidates, and we will need to test the significance of their parameters.

After using the coeftest() function, we perform a z-test of the coefficients and obtain the following estimates, along with their p-values for Model 2 and 3:

Table 2: Coefficient Estimates and P-values for Model 2 & 3



Coefficient	Estimate	Standard Error	P-Value
Model 2			
ma1	-0.4073	0.0518	0
ma2	-0.1294	0.0482	0.0073
sar1	0.073	0.0645	0.2578
sma1	-0.8792	0.0376	0
Model 3			
ma1	-0.4057	0.0518	0
ma2	-0.1285	0.0481	0.0076
sma1	-0.8578	0.0354	0

The estimates and p-values for both models are almost identical, with the exception of one coefficient: the seasonal autoregressive order. Since the p-value for it is way above our significance threshold, we cannot claim that the coefficient is non-zero. Due to this, we have to concede Model 2 to Model 3 and have it as our final model. For the non-seasonal moving average coefficients and the seasonal moving average coefficient, all three p-values are significant at the 0.01 alpha level. Hence, we can conclude that these coefficients are non-zero.

Therefore, the final model is Model 3, and its final equation is:

$$(1-B)(1-B^{12})x_t = (1-0.4057_{0.0518}B - 0.1285_{0.0481}B^2)(1-0.8578_{0.0354}B^{12})w_t$$

where the left side represents the product of the seasonal and first difference, and the right hand side is the product of the non-seasonal MA(2) model and the seasonal MA(1) with period 12.

We will now forecast the data into the future 12 months, comparing it to the last 12 observations we had omitted.

Forecasting and Spectral Analysis

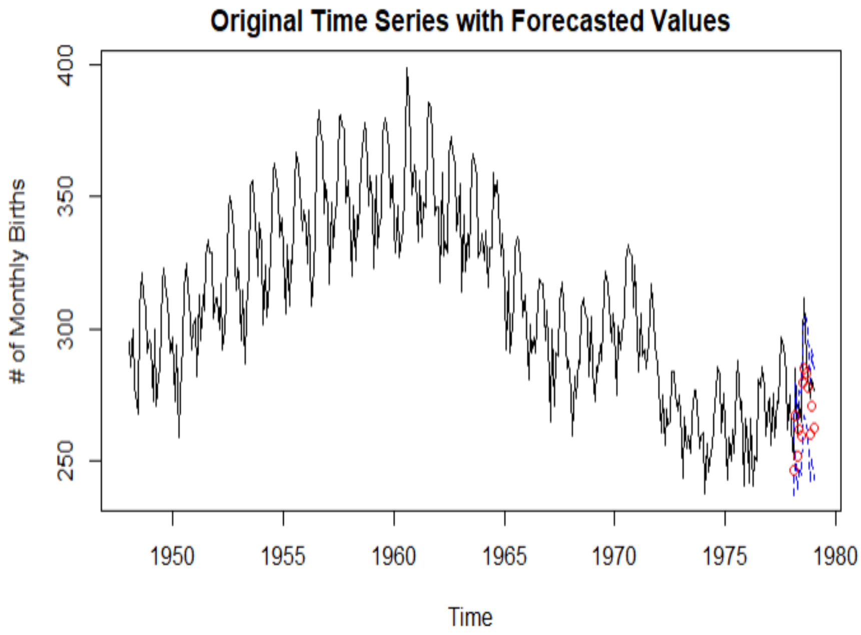
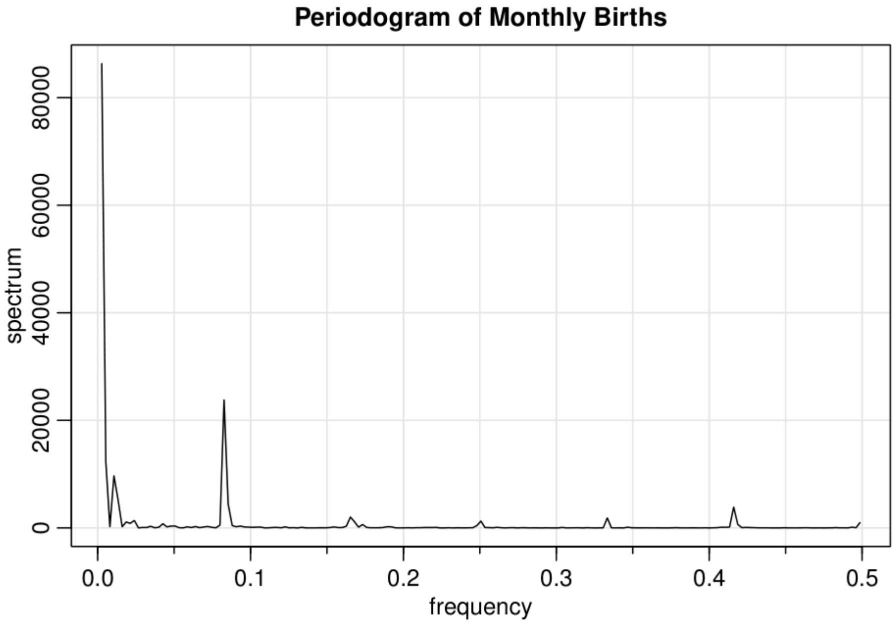


Table 3: Prediction Intervals and Forecasted Values

Lower Bound	Upper Bound	Observed Value
237	257	251
255	281	285
240	265	260
248	277	272
245	275	265
263	298	296
267	305	312
265	304	289
259	298	282
242	280	274
251	292	281
243	284	277

It looks like most of the observations from our testing data fall into their appropriate prediction interval. With 95% confidence, these values will have fallen in these ranges.

Lastly, we will perform a spectral analysis to identify the first three predominant periods and obtain the respective confidence intervals.



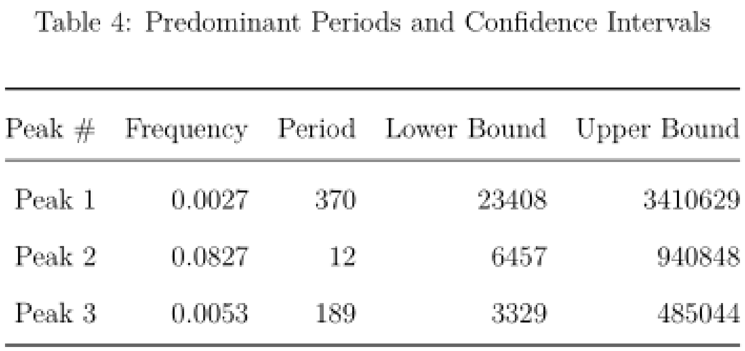


Table 4: Predominant Periods and Confidence Intervals

Peak #	Frequency	Period	Lower Bound	Upper Bound
Peak 1	0.0027	370	23408	3410629
Peak 2	0.0827	12	6457	940848
Peak 3	0.0053	189	3329	485044

If we look at the periods, we must divide them by 12 since we are dealing with monthly observations. For peaks 1 and 3, it looks like there is a dominant periodicity every 15 and 30 years. Peak 2 suggests a smaller frequency every month, which may have to do with how the data was collected (perhaps a spike of observations was seen at the end of every month). For every dominant frequency, we are 95% confident that its corresponding spectrum lays within that interval. However, these confidence intervals are too wide to actually make use of, and so their significance is questionable.

Discussion and Conclusion

In all, the original intention of this report was to analyze the U.S monthly births using data from the past to fit the most optimal time series model and use it to forecast/predict values in the future. This was done by detrending, deseasoning, and differencing the data in order to produce a valid time series. Furthermore, we were able to forecast 10 of the 12 observations, where birth rates have considerably decreased in the later years, matching the original hypothesis and our testing data. One of the main limitations is the seasonal ARIMA model choice. These models were instinctively chosen with visual inspection of the ACF/PACF, and so not every possible model combination was considered for the most optimal fit. In addition, we could not correctly forecast an entire year with our best model. Hence, there might be some extraneous factors that the model was not able to capture. Overall, we have done a solid job, and it was very interesting to analyze the fluctuation of birth rates in the past years, as these results coincide with the low birth rates seen today. (**last page is the bibliography.**)

Bibliography

"US Birth Rate Falls 4% to Its Lowest Point Ever." *BBC News*, May 5, 2021, sec. US & Canada. https://www.bbc.com/news/world-us-canada-57003722.

8.1 Stationarity and Differencing | Forecasting: Principles and Practice (2nd Ed). Accessed December 16, 2021. https://Otexts.com/fpp2/.

Editors, History com. "Baby Boomers." HISTORY. Accessed December 16, 2021. https://www.history.com/topics/1960s/baby-boomers-1.

PRB. "Why Is the U.S. Birth Rate Declining?" Accessed December 16, 2021. https://www.prb.org/resources/why-is-the-u-s-birth-rate-declining/.

Shumway, Robert H., and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-52452-8.

Zach. "Augmented Dickey-Fuller Test in R (With Example)." *Statology* (blog), May 25, 2021. https://www.statology.org/dickey-fuller-test-in-r/.

Zach. "How to Perform a Box-Cox Transformation in R (With Examples)." *Statology* (blog), October 13, 2020. https://www.statology.org/box-cox-transformation-in-r/.