Time series analysis of Finland monthly birth during 2010-2020

Monitoring and predicting population dynamics is very important for policy making. In this study, I analyzed the monthly birth data of Finland during 2010-2020. This monthly birth shows a unimodal annual cycle with a decreasing trend over the last ten years. I found that this time series can be fitted by a SARIMA model. This model can predict the testing data with high accuracy. It also indicates monthly birth in Finland will keep decreasing in 2020-2021.

Introduction

Different from many animals, human can give birth throughout the year. However, various studies showed that human reproduction has a pronounced annual rhythm. The waveform of the annual rhythm (e.g., unimodal or bimodal) varies among geographical regions and persists for many years. The amplitude or phase of the rhythm can also change over time [1]. It is an interesting question to investigate which environmental or social factors affect this reproduction rhythm [2]. What’s more, population dynamics has a strong effect on economic growth [3], monitoring the population dynamics can help the government better prepare potential population issues.

The dataset I use is from Statistics Finland, the national statistical institution in Finland [4]. The dataset describes the monthly birth data during 2010-2020. It can be used as an example for the reproduction annual rhythm in high-income European countries. I hoped to build a time series model to describe this annual rhythm and predict the future trend. I found that there is a strong 12-month cycle in monthly birth: the birth number is highest around summer and lower at the beginning or the end of each year. There is also a significant decreasing trend through the last ten years. Therefore, SARIMA models were used to fit the data during 2010-2019. I chose the best of them according to the significance of coefficients and AICCs. I found that the model can accurately predict the data from 2019 to 2020. It also predicts that the monthly birth in 2020-2021 will keep decreasing. All analysis was done in R.

Results

Data exploration

In figure 1, we can see that there is a 12-month cycle for monthly birth. The birth counts are highest in the middle of each year. There is also a strong decreasing trend over the last ten years. The average number decreases for more than 1000 from 2010 to 2020. I used the data during 2010-2019 as the training data, and the data during 2019-2020 as the testing data.

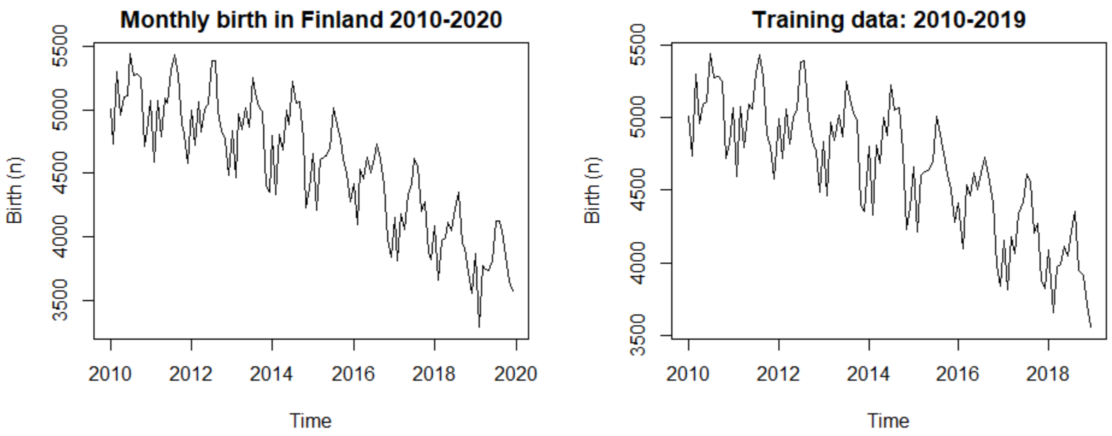

Figure 1. Left: Monthly birth in Finland during 2010-2020. Right: Training data used for model fitting.

Data transformation

The time series does not show big changes in variance over time. Therefore, no data transformation was used. I differenced the data at lag 1 to remove the trend (Figure 2A), then differenced at lag 12 to remove the seasonality (Figure 2B). Data variance decreases from 199158 to 74862 to 30512. The resulting time series looks stationary. The data distribution also shows nearly Gaussian (Figure 2C).

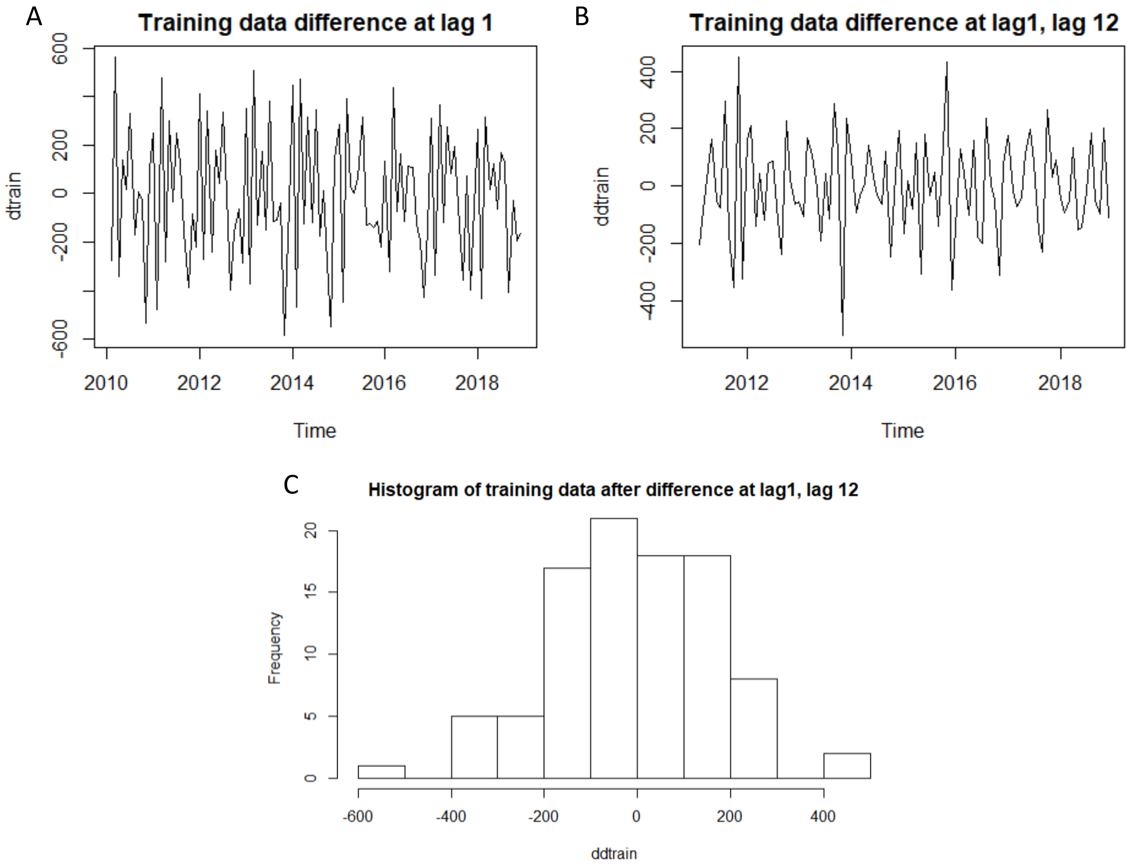

Figure 2. **A.** Training data after difference at lag 1. **B.** Training data after difference at lag 1 and lag12. **C.** Histogram of differenced training data.

Model fitting

Since the original data shows a 12-month cycle and a linear trend, I decided to use SARIMA model and chose D=1, d=1. To choose for the other parameters, I analyzed the ACFs and PACFs for the differenced time series.

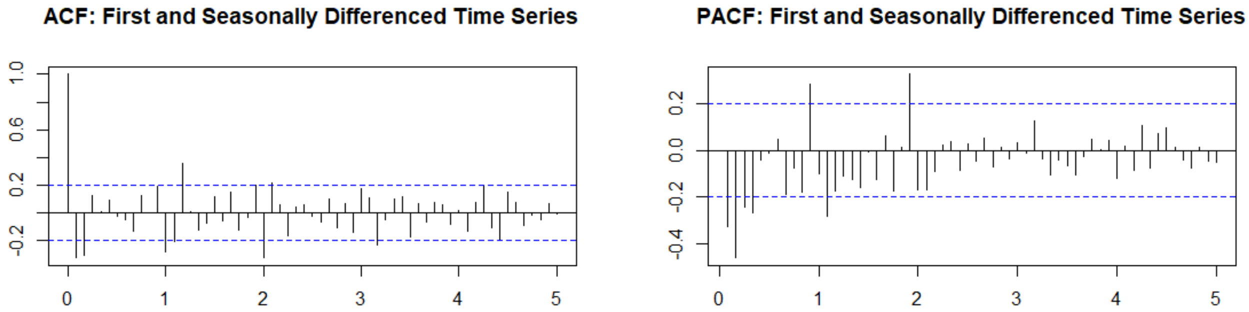

Figure 3. Left: ACFs of differenced training data. Right: PACFs of differenced training data.

From the ACFs result, we can see that the ACFs at lag 12 and lag 24 are outside confidence intervals. ACFs at lag 1 and lag 2 are also outside confidence intervals. It indicates Q=2 and q=2 for the SARIMA model. From the PACFs result, we can see that the PACFs around lag 12 and lag 24 are outside confidence intervals. PACFs at lag 1-4 are also outside confidence intervals. It indicates P=2 and p=4 for the SARIMA model. I differenced the original data at lag 1 and lag 12, therefore, D=1, d=1 and s=12. I fitted the models using the following parameters and compare their AICCs. The first model could not be fitted because infinite values were generated during optimization.

P	p	Q	q	AICCs
2	4	2	2	Could not be fitted
2	2	2	2	1181.03
1	4	2	2	1183.73
1	4	1	2	1188.75
1	4	2	0	1182.49

Table 1. AICCs of SARIMA models fitted by different parameters.

Table 1 shows that the AICCs of the second and the last models are small. Therefore, I checked these two models to choose the better one from them.

SARIMA (P=2, p=2, D=1, d=1, Q=2, q=2, s=12) σ^2=9972								
	ar1	ar2	ma1	ma2	sar1	sar2	sma1	sma2
Coefficients	-0.5056	-0.2457	-0.2335	-0.2568	-0.934	-0.3501	0.3397	-0.5273
S.E.	0.3602	0.1553	0.3492	0.3055	0.2291	0.2064	0.5469	0.5039

Table 2. Coefficients and their standard errors of SARIMA (P=2, p=2, D=1, d=1, Q=2, q=2, s=12)

SARIMA (P=1, p=4, D=1, d=1, Q=2, q=0, s=12) σ^2=9936							
	ar1	ar2	ar3	ar4	sar1	sma1	sma2
Coefficients	-0.6864	-0.6594	-0.3306	-0.1806	-0.693	0.166	-0.8334
S.E.	0.1038	0.1207	0.1196	0.1029	0.124	0.2516	0.2294

Table 3. Coefficients and their standard errors of SARIMA (P=1, p=4, D=1, d=1, Q=2, q=0, s=12)

In table 2, the confidence intervals of many coefficients include 0, indicates they are not significant. Most coefficients in table 3 are significant. Therefore, I chose this SARIMA $(p = 4, d = 1, q = 0) \times (P = 1, D = 1, Q = 2)_{s=12}$ model as the final model to test. The confidence intervals of ar4 and sma1 coefficients include 0. I assigned sma1 to 0 and refitted the model using p=3. The resulting model has lower AICC (1182.15). And all its coefficients are significant (table 4). The model parameters are little different from what are suggested by ACFs/PACFs of differenced time series. This model for monthly birth X_t can be written as:

$$(1+0.647B+0.558B^2+0.219B^3)(1+0.66B^{12})(1-B)(1-B^{12})X_t=(1-0.999B^{24})Z_t \qquad Z_t \sim WN\ (0,\ 8984)$$

SARIMA (P=1, p=3, D=1, d=1, Q=2, q=0), s=12 σ^2=8984						
	ar1	ar2	ar3	sar1	sma1	sma2
Coefficients	-0.6465	-0.5578	-0.2192	-0.6604	0	-0.9988
S.E.	0.103	0.1064	0.102	0.1171	0	0.3444

Table 4. Coefficients and their standard errors of the final SARIMA (P=1, p=3, D=1, d=1, Q=2, q=0, s=12) after assigning the sma1 coefficient to 0.

I then checked whether the AR part of this SARIMA model is stationary. From figure 4, we can see that all roots for AR part are outside the unit circle. Therefore, the AR part is stationary.

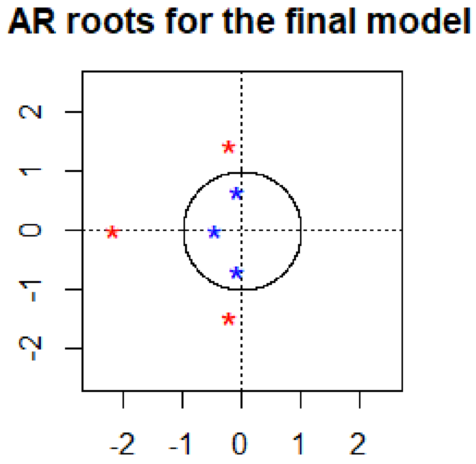

Figure 4. The roots of AR characteristic polynomial. Red stars stand for the roots, blue stars stand for the inverse of the roots. X-axis stands for the real number part of the roots. Y-axis stands for the imaginary number part of the roots.

Diagnostic checking

Next, I performed analysis on model residuals to see whether the model is adequate. The residuals mean is -9.62, which is small compared to the range of the data. The fitted residuals do not have trend or seasonality, and do not show change of variance over time. Almost all ACFs and PACFs are within the confidence intervals (figure 5). Therefore, the fitted residuals resemble white noise.

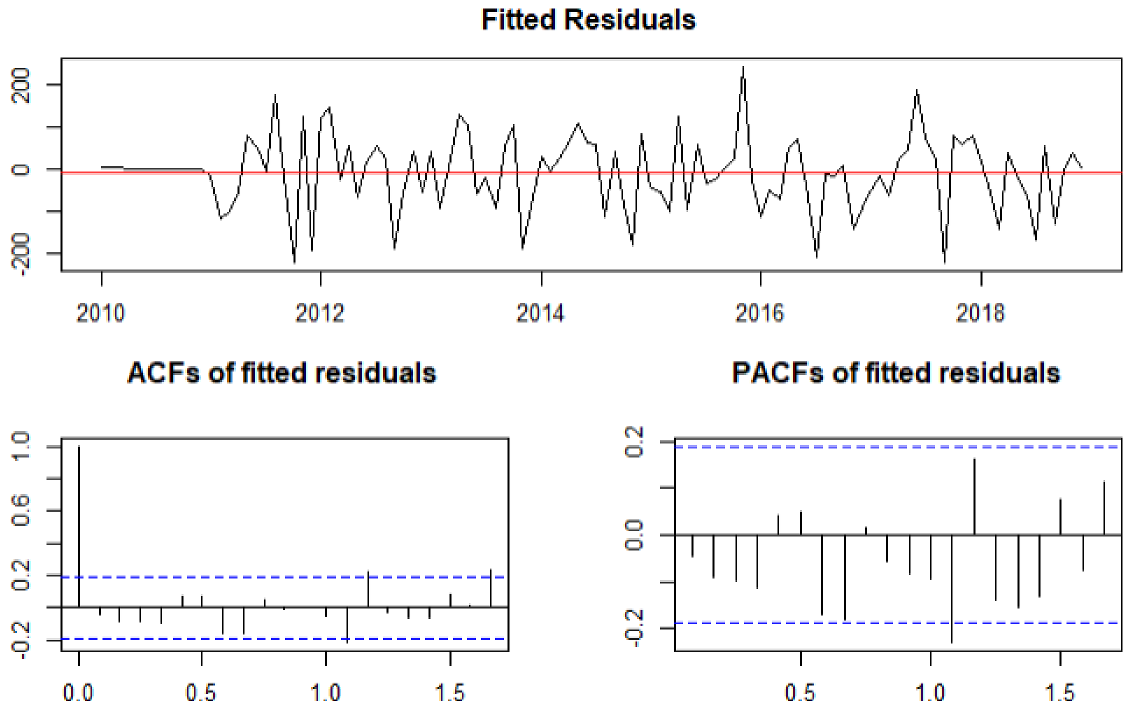

Figure 5. The model residuals and their ACFs and PACFs.

Portmanteau tests were also used to check the correlation between residuals or the squared residuals. Box-Pierce, Ljung-Box and McLeod-Li tests were used. For Box-Pierce and Ljung-Box tests, I chose the degree of freedom as 5 (10-5). For McLeod-Li test, I chose the degree of freedom as 10. None of the tests show significance. This indicates residuals or the squared residuals are not correlated. The fitted residuals resemble white noise.

Test	Box-Pierce	Ljung-Box	McLeod-Li		Shapiro-Wilk
P-value	0.08691	0.06539	0.9415		0.169

Table 5. The P-values of Box-Pierce, Ljung-Box and Shapiro-Wilk test on model residuals, and McLeod-Li test on the squared residuals.

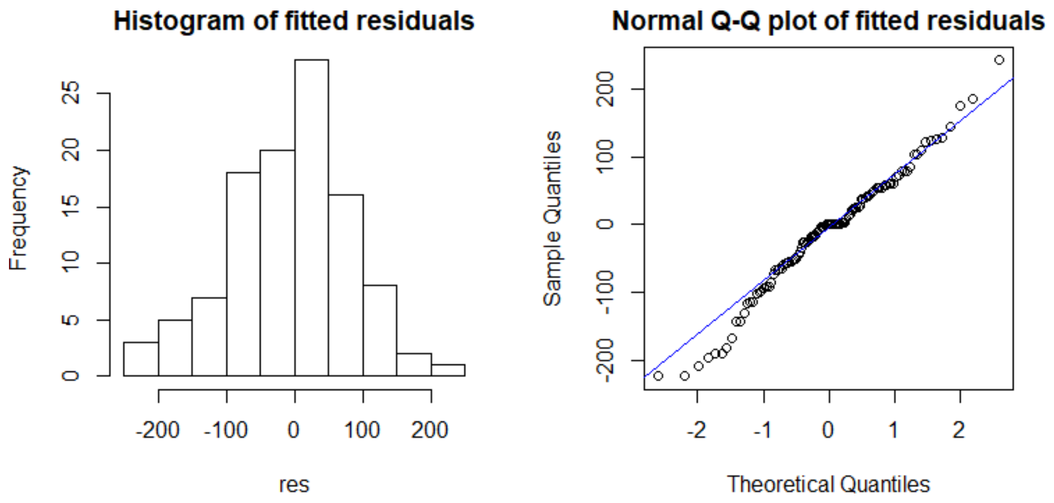

Figure 6. The histogram and Q-Q plot of residuals.

Last, I checked the distribution of fitted residuals. The histogram and Q-Q plot resemble Gaussian distribution but with longer tails. Shapiro-Wilk test doesn’t show significant difference from normal distribution. The residuals are approximately Gaussian.

In summary, our SARIMA model passes all diagnostic checking and is adequate for the training data.

Model forecasting

To further investigate whether this SARIMA model can predict the further trend of monthly birth in Finland. I made predictions for monthly birth in 2019-2021. The predictions in 2019-2020 are confirmed by the testing data: all ground truths are within the 95% confidence intervals. The model also predicts that the monthly birth will keep decreasing during 2020-2021.

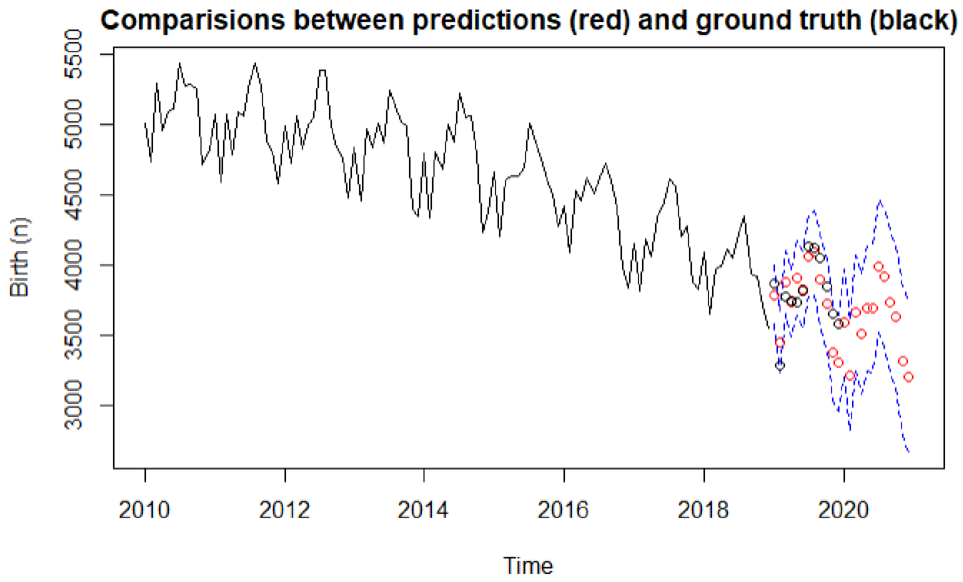

Figure 7. The predictions of Finland monthly birth during 2019-2021. Red circles stand for the predictions and blue dashed line stands for their 95% confidence intervals. Black circles stand for the testing data during 2019-2020.

Conclusion

In this project, I analyzed Finland monthly birth data during 2010-2020. I found that the time series from 2010 to 2019 can be fitted by a SARIMA model:

$$(1+0.647B+0.558B^2+0.219B^3)(1+0.66B^{12})(1-B)(1-B^{12})X_t=(1-0.999B^{24})Z_t \qquad Z_t \sim WN\ (0,\ 8984)$$

This model can predict the testing data during 2019-2020 with high accuracy. It also indicates monthly birth in Finland will keep decreasing in 2020-2021. Therefore, the Finland government should take measures to increase birth rate and prepare for the population aging problem in the future.

Reference

1. Roenneberg, Till, and Jürgen Aschof. "Annual rhythm of human reproduction: I. Biology, sociology, or both?." Journal of biological rhythms 5.3 (1990): 195-216.

2. Roenneberg, Till, and Jürgen Aschoff. "Annual rhythm of human reproduction: II. Environmental correlations." Journal of Biological Rhythms 5.3 (1990): 217-239.

3. Peterson, E. Wesley F. "The role of population in economic growth." SAGE Open 7.4 (2017): 2158244017736094.

4. 11 II – Vital statistics by month.
http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/StatFin__vrm__vamuu/