Prediction of Birth Rate of American Samoa

Xusen Shi*

College of Letter and Science, University of California Santa Barbara, 93106, Isla Vista, USA

*Corresponding author. Email: xusenshi@ucsb.edu

Keywords: American Samoa, SIMA model, prediction, birth rate.

Abstract: The main purpose of this paper is to find the pattern of the number of births of American Samoa against time and make a prediction of the number of births of American Samoa in the next few years to see will the number of births in the future be steady. Samoa is an unorganized American territory that has over 1000 years of history. In this project, a model of the number of births is created depending on the data set of the number of births in the past 25 years. The model is created for pattern and prediction. A SIMA model is fitted, and forecasting is done. The article finds that the number of births in the future will be steady. There will not be a sharp increase or decrease in the number of births.

1. Introduction

American Samoa is an unorganized territory of America. It is located in the South Pacific and is a great natural harbor [1-2]. One thousand thousand years ago, people started to live there, and it has a really interesting history. In the late 19th century, England, Germany, and America had a serious conflict of the belonging of Samoa [3]. In 1899, America and Germany signed a contract to split Samoa into halves. Currently, west Samoa is an independent country, and east Samoa belongs to American [4-6]. It will be interesting to study the population of a secret place with a long and dramatic history. The number of births each year is an important component of the population, and will the number of births each year be steady in the future is also important. Thus, this project will concentrate on the pattern of the number of births in Samoa in the past 25 years and forecast the number of births in Samoa in the next few years.

The data set, Seasonal Variation in Birth, is from Kaggle. It includes data on the number of births every month for 135 countries. The number of births in Samoa in the past 25 years is selected from this data set. There are a total of 300 observations.

In order to find the pattern and do the forecast, a time series model needs to be created. In order to create a model, the raw data need to be modified. The raw data need to be differenced at certain lags to eliminate trend and seasonality if they exist. Transformation needs to be performed if the variance is not constant. In order to perform the apparent transformation, the lambda of Box-Cox transformation needs to be found. After the time series is stationary, the preliminary model can be identified by the graphs of ACF and PACF [7-9]. The model with the lowest AICC will be chosen among all the preliminary models.

After the model is fitted, the poly roots need to be found in order to check stationary and invertibility. Also, diagnostic tests need to be performed: Box-Pierce test, Ljung-Box test, and Shapiro test [10-11]. If all the tests are passed, forecasting can begin. Otherwise, the model needs to be adjusted according to the test result.

In order to create the model, the variance of the time series needs to be constant, or a transformation needs to be performed. Second, in order to have a SARIMA or ARIMA model, the time series should be stationary and invertible. For a stationary time series, there should not be any trend or seasonality. Thus, trends and seasonality need to be detected if they exist. A linear model of time series can be created and plotted with the time series and mean of time series to detect the trend. The graphs of ACF and PACF can detect the seasonality. If trend and seasonality exist, the time series needs to be differenced at certain lags. After differencing, the time series, ACF, and PACF should be plotted again to check if the time series is stationary. If the time series is stationary, the preliminary models can be identified by the graphs of ACF and PACF. Then, all the potential models will be fitted, and the model with the lowest AICC will be chosen.

After choosing the model, some diagnostic tests are performed, Box-Pierce test, Ljung-Box test, and Shapiro test. Poly roots need to be found as well to ensure invertibility and stationary. If all the tests are passed, forecasting can be progressed. Finally, a SARIMA model is created, and the forecast results show that in the next few years, the number of births of Samoa will be steady and will not have a huge increase or decrease.

The analysis will start with a raw data observation. The purpose of this step is to detect some obvious patterns. The time series of raw data will be plotted at this step. Then, transformation is performed in order to eliminate the problem of nonconstant variance. After variance is stabilized, variance and trend need to be eliminated as well because a stationary time series is needed for a SARIMA model. Finally, the model will be fitted, and the model with the highest AICC will be selected. After the model selection, a diagnostic test will be performed to ensure the model is suitable. At last, the prediction of the birth rate will be calculated

2. Raw Data Observation

The data set, Seasonal Variation in Births, is from Kaggle.com. It includes a number of births every month in 135 countries. Data used in this paper only focused on the number of births of Samoa, so the number of births from 1994 to 2019 if Samoa is selected from the data set.

The time series of the raw data is plotted as figure 1 with the mean and the trend. Apparently, there is an upper trend. The problem of variance being not constant and seasonality is not obvious. Further analysis is needed.

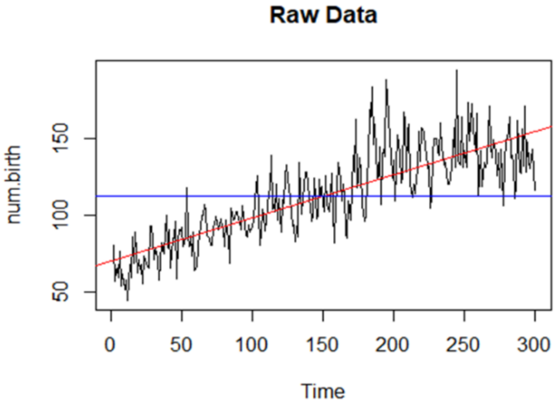

Figure 1 Time Series of Raw Data

In order to fit a Seasonal Autoregressive Integrated Moving Average model, the time series need to have no trend, no seasonality, and constant variance. According to the graph, an upper trend is apparent. However, by seeing the graph, seasonality and variance cannot be determined, so further analysis is needed.

The data set is partitioned into a training set and testing set. The training set has 270 observations, and the testing set contains 30 observations. The ratio is 9:1. This ratio is chosen because the model will be more accurate with more observations in the training sets. The time series of the training set is plotted as figure 2 called **U_t.**

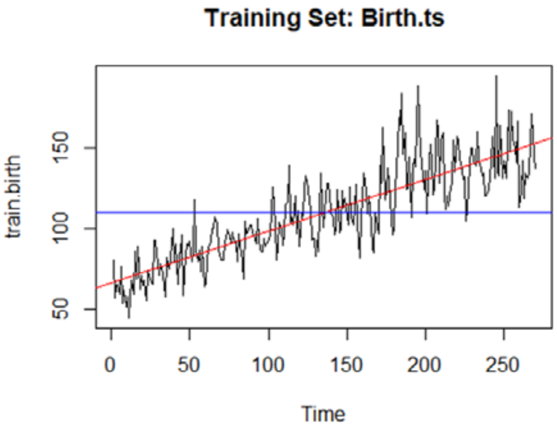

Figure 2 Time Series of Training Set It is similar to the original data set.

3. Transformations

If the variance is not constant, the transformation is needed in order to find the proper transformation of the data set. The Box-Cox transformation of the training set is performed. The lambda = 0.3838 is calculated from the transformation, which means that the transformation is needed. Figure 3 shows that the 0.5 and ⅓ is with the confidence interval, so the potential transformations are square root, cubic root, and Box-Cox.

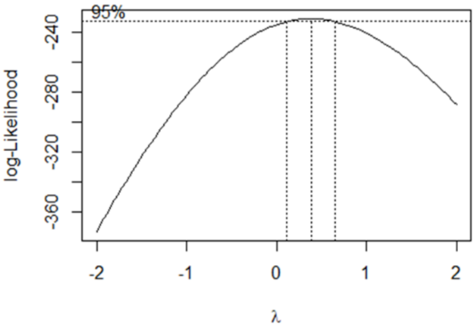

Figure 3 Box-Cox Transformation of Ut

After all three transformations are performed, all three transformations' time series are plotted. Figure 4 is the graph of Box-Cox transformation. Figure 5 is the graph of square root transformation. Figure 6 is the graph of The graphs don't show obvious differences.

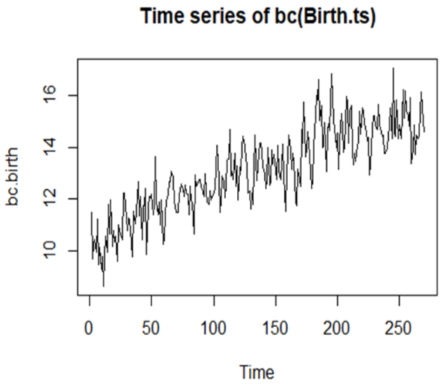

Figure 4 Box-Cox Transformation of Training Set

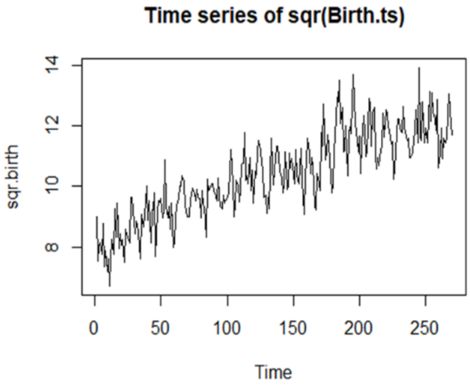

Figure 5 Square Root Transformation of Training Set

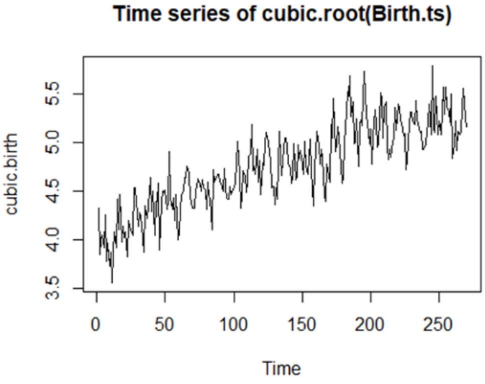

Figure 6 Cubic Root Transformation of Training Set

In order to find the most appropriate transformation, the variances of all transformations are calculated, and the histograms of all transformations are plotted. Figure 7 is the histogram of Box-Cox transformation. Figure 8 is the histogram of square root transformation. Figure 9 is the histogram of cubic root transformation. The cubic root transformation is left-skewed, so we don't want this transformation, although it has the least variance. Compared to the Box-Cox transformation and the square root transformation, the square root transformation will be more symmetric. Moreover, the variance of the training set is 864.9628, the variance of Box-Cox transformation is 2.71715, and variance of square root transformation is 2.00266, and the variance of cubic root transformation is 0.1888254. The square root transformation has the second least variance and is the most symmetric. Thus, I choose the square root transformation called **(U_t)^½.**

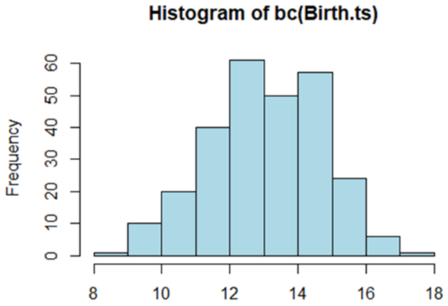

Figure 7 Histogram of Box-Cox Transformation

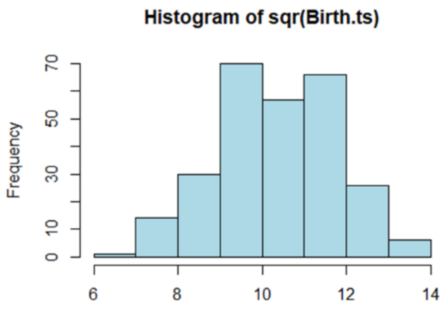

Figure 8 Histogram of Square Root Transformation

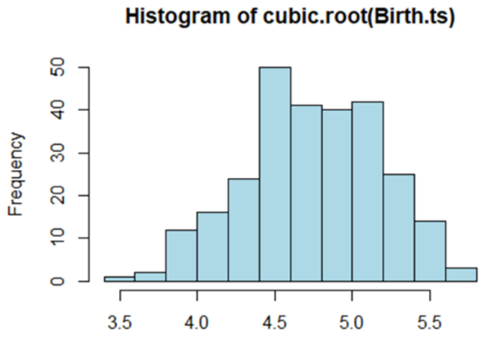

Figure 9 Histogram of Cubic Root Transformation

The time series is decomposed to detect trend and seasonality. The decomposition shows that the time series have trend and seasonality. Figure 10 shows the decomposition.

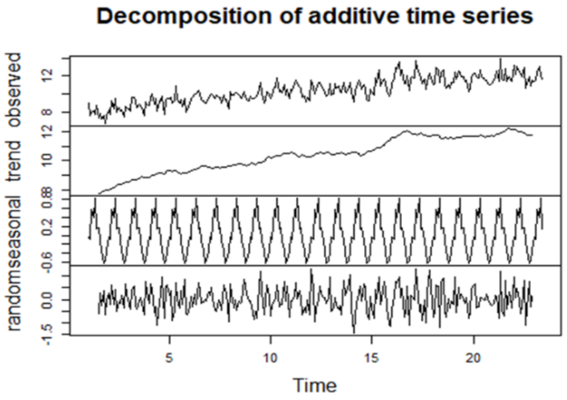

Figure 10 Decomposition of Additive Time Series

The figure shows that the time series has a trend and seasonality. Also, figure 11, the graph of the time series shows an obvious upper trend, and figure 12. the graph of ACF of the (U_t)^½ shows apparent seasonality. Thus the time series is not stationary, and it needs to be differenced.

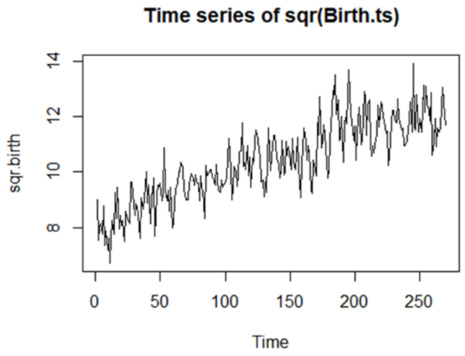

Figure 11 Time Series of Training Set

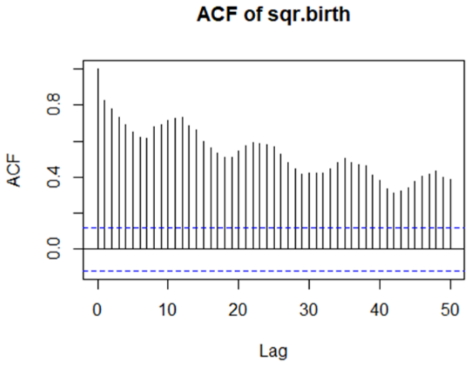

Figure 12. ACF of Training Set

ACF is the autocorrelation function of the time series

In order to eliminate the seasonality and trend, the time series is differenced at lag 1, which is called **∇1(Uₜ)^½.** This time series and its ACF are plotted as graph 4.4 and graph 4.5. Both graphs show that the seasonality no longer exists, but there is still a lower trend. Moreover, after differencing at lag 1, the variance decreases

4. Method

4.1 Time Series Introduction

A Seasonal Autoregressive Integrated Moving Average Model will be fitted. The formula of this model will be

$$\varphi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t \qquad (1)$$

B means the backward transformation. Phi is the coefficient of the dependent variable, and theta is the coefficient of white noise. Z means the variable

4.2 Identify Preliminary Models

Figure 13 is the ACF of the time series, and figure 14 is the PACF of the time series. PACF means the partial autocorrelation function of time series.

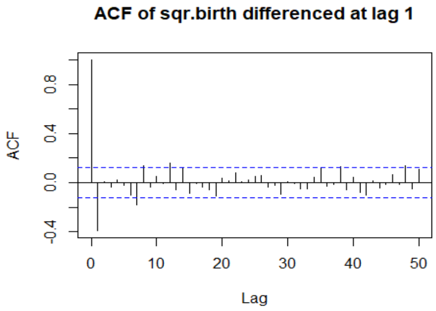

Figure 13 ACF

Then, the potential model will be

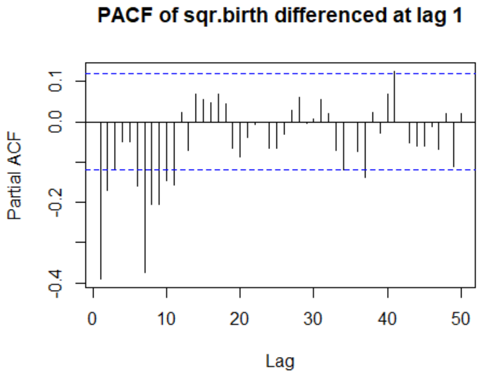

Figure 14 PACF

Table 1 Potential Models

P	**D**	Q	p	d	q	AICC
1	1	1	0	0	1	565.7098
0	1	1	0	0	1	576.5293
7	1	1	0	0	1	539.116

Finally, the fitted model will be

$$\nabla_1(U_t)^{1/2} = (1-0.1566B^5-0.2596B^6-0.2518B^7)(1-B)X_t + (1-0.7611B)(1+0.1027B^{12})Z_t \quad (1)$$

5. Diagnostic Checking

Diagnostic checking needs to be performed to ensure the fitted model is appropriate. First, all the absolute value of the coefficient is smaller than one, so no unit-roots will exist. The graph of all roots are plotted as graph 15

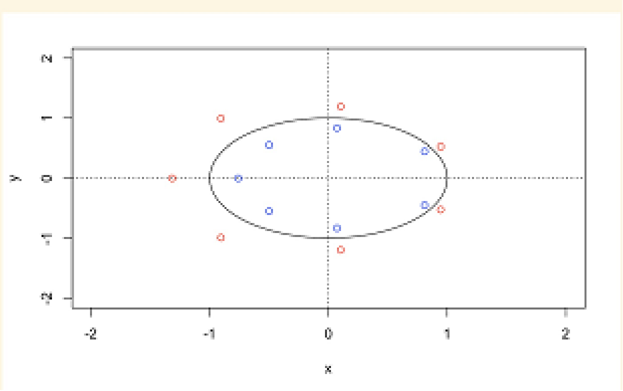

Figure 15 Unit Roots of Fitted model

The ellipse in Figure 15 is a unit circle. The red points are the unit roots, and all of them are outside the unit root, which means that we have a stationary and invertible model.

Then, the Box-Pierce test, Ljung-Box test, and the Mcleod-Li test are passed. All the p-values are greater than 0.05. The test result is shown as figure 16

Table 2 Test Result

Test name	Degree of Freedom	p-value
Box-Pierce Test	11	0.658
Box-Ljung Test	11	0.6283
McLeod Test	16	0.8181

In conclusion, the model is suitable.

6. Prediction

For forecasting the transformed time series, the figure of prediction of the training set is plotted with confidence interval as Figure 16. The pattern matches the original data well, which means that it is a successful model for birth rate prediction.

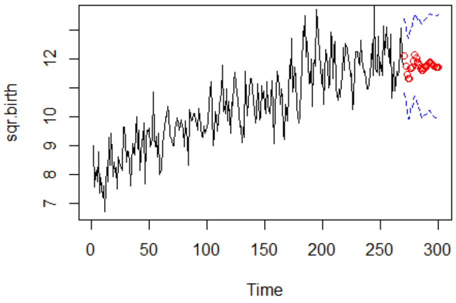

Figure 16 Prediction of Training Set

Then, figure 17 is the prediction of the future birth rate of the next 30 months. There is no sharp increase or decrease in birth rate in the next 30 months which means the population in the future will be steady for American Samoa.

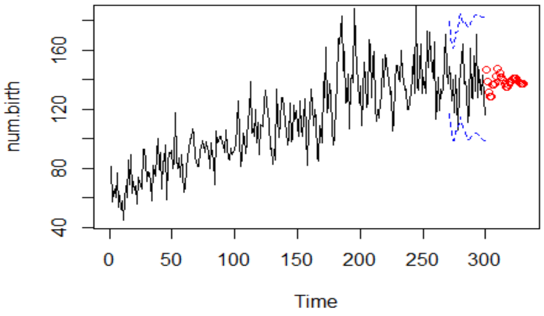

Figure 17 Prediction of Original Data

7. Conclusion

The purpose of this paper is to predict the future birth rate of American Samoa. American Samoa is an unorganized colony of Americans. It has a long and interesting history. It belonged to several countries before becoming the colony of America. Its special location attracts many countries to

conquer this place. Thus, there is a cultural shock and exchange at this place. In order to forecast the future birth rate of American Samoa, a SARIMA model is built, which is a time series model.
Using the model for prediction, we find out that the birth rate of this place will not fluctuate a lot. It will stay within a stable range of about 130 thousand a year. Thus, the population of this place will be stable in the next few years. A stable population implies a stable social situation. It is glad to see a place with complicated history can be stable.

However, the prediction can be improved in servals. First, the data set is not large enough. In the step of prediction, we can see that the confidence interval is wide. Larger sample size can narrow down the confidence interval. Moreover, other models can be used for prediction. They may be more accurate. Other features may be contributed to the change of birth rate besides birth rate in the past and time. Another model can include all other influential features to make a more accurate model. Thus, a larger sample size with more features is needed for further development, and a more accurate model and algorithm is needed

References

[1] Lu, W., Li, J., Wang, J., & Qin, L. (2021). A CNN-BiLSTM-AM method for stock price prediction. Neural Computing and Applications, 33(10), 4741-4753.

[2] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. The Journal of finance and data science, 4(3), 183-201.

[3] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 205-208). IEEE.

[4] Hu, Z., Zhao, Y., & Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. Applied System Innovation, 4(1), 9.

[5] Ferdiansyah, F., Othman, S. H., Radzi, R. Z. R. M., Stiawan, D., Sazaki, Y., & Ependi, U. (2019, October). A lstm-method for bitcoin price prediction: A case study yahoo finance stock market. In 2019 International Conference on Electrical Engineering and Computer Science (ICECOS) (pp. 206-210). IEEE.

[6] Chohan, U. W. (2021). Counter-hegemonic finance: The gamestop short squeeze. Available at SSRN.

[7] Zhang, D. (2018). Energy finance: background, concept, and recent developments. Emerging Markets Finance and Trade, 54(8), 1687-1692.

[8] Veloso, M., Balch, T., Borrajo, D., Reddy, P., & Shah, S. (2021). Artificial intelligence research in finance: discussion and examples. Oxford Review of Economic Policy, 37(3), 564-584.

[9] Li, W., Cheng, Y., & Fang, Q. (2020). Forecast on silver futures linked with structural breaks and day-of-the-week effect. The North American Journal of Economics and Finance, 53, 101192.

[10] Wei, Y., Liang, C., Li, Y., Zhang, X., & Wei, G. (2020). Can CBOE gold and silver implied volatility help to forecast gold futures volatility in China? Evidence based on HAR and Ridge regression models. Finance Research Letters, 35, 101287.

[11] Hu, W. (2021). Volatility Forecasting of China Silver Futures: the Contributions of Chinese Investor Sentiment and CBOE Gold and Silver ETF Volatility Indices. In E3S Web of Conferences (Vol. 253, p. 02023). EDP Sciences.