Is the US Film Industry at Risk? A Time Series Analysis for Recent Domestic Box Office Data

Xiru Lyu

May 1, 2018

Summary: This paper attempts to develop a time series model that captures the recent movement of the US domestic box office and makes good predictions of its future performance. Using monthly gross box office data from January 2008 to March 2018 downloaded from the reporting service boxofficemojo.com, several ARIMA and SARIMA models were developed and evaluated. With a total of 123 observations, the whole dataset was divided into the training set, which contains monthly gross box office from January 2008 to December 2015, and the test set, which is made up by monthly data from January 2016 to March 2018. Following transformation of the time series and examination of potential ARIMA and SARIMA models, comparisons among forecasts by ARMA models and by non-ARMA method were carried out for choosing the one with highest forecast accuracy. The resulting model, SARIMA$(1,1,1) \times (0,1,1)_{12}$ with parameters $\phi = 0.0257, \theta = -0.999, \Theta = -0.9977$ is considered to both fits the data and makes reliable forecasts. Many recent articles point out that the US film industry is experiencing its downturn, and thus sectors related to the industry are expected to adjust for its future movements by utilizing results of this study. The finalized SARIMA model predicts that the overall trend for domestic box office would become flattened after March 2018 compared with the previous upward trend, suggesting a slowdown of the US film industry. However, the stability of the finalized model is still in question since the rising ticket price and lowering in domestic ticket sales might undermine its predictive power.

1 Introduction

Many people choose to watch movies during their free time for entertainment. According to Statista (2018), the US is the third largest film market in the world, with about 13% of Americans go to see movies about one every month, while about 7% of Americans visit movies theaters several times a month. Besides, Hollywood continues to play a dominant role in the world film market as many popular movies are produced by Hollywood. Due to the importance of the US film industry, proper functioning of many sectors, such as film investment companies and production companies, depend heavily on its performance. Thus, it is important to analyze the recent movement of the industry as well as to predict its future. This study is devoted to analyzing domestic monthly gross box office data to develop a time series model that both fits the data and produces good predictions. Several ARIMA and SARIMA models were created from subset selection and plot identification, and subsequent model evaluation was performed. The best model was chosen through forecast comparisons among ARMA models and non-ARMA method. Context of the research topic as well as detailed Box-Jerkins analysis for selecting the best fitted model are given in the following sections.

2 Background

Many recent articles express pessimistic view of the US film industry. Plaugic (2018) describes the decline of the industry as it is threatened by the surge of ticket price and increasing competition from streaming services like Netflix, Amazon and HBO GO. Bilton (2017) points out that stimulating original contents by Netflix hurt the show business as well as the traditional film industry. As mentioned in the introduction, the US movie industry has significant influence to many sectors, and thus its underperformance would have wide impacts. Thus, it is necessary to forecast and assess the future of the US film industry so that related sectors can adjust to changes accordingly.

The main focus of the study is the domestic box office data since it is the product of the show business, while the show business directly relates to the film industry. In order to better capture the trend of the box office, the series is set to cover the most recent 10-year-period, from January 2008 to March 2018. Box office data used in this study were downloaded using R package "boxoffice" from the leading online box-office reporting service boxofficemojo.com, owned and operated by the movie website IMDb. The raw dataset contains movie information and daily box office data for movies on view in US theaters from 2008-01-01 to 2018-03-31. No missing values or outliers were detected in the dataset. During the data cleaning process, only two variables – every movie's daily gross box office and date that the box office number was collected were kept within the dataset. Then the box office data was summed up by month and year to produce monthly data, and eventually a time series of 123 observations from January 2008 to March 2018 with a frequency of 12 was created. In addition, gross box office was adjusted for inflation using inflation rate found in Federal Reserve Bank of St. Louis's database, with the price index of 2008 setting to be 100 and indices of other years calculated based on yearly inflation rate in relation to 2008's. Since there is no annual inflation rate for 2018 yet, 2018 box office numbers were adjusted using the price index for 2017. Finally, box office data were rescaled in millions of dollars.

3 Modeling & Diagnostics

3.1 Descriptive Analysis

For the purpose of assessing forecast accuracy, the dataset was split into training and test sets. The training set contains 96 observations of monthly gross box office from January 2008 to December 2015. The rest of the data were put in the test set.

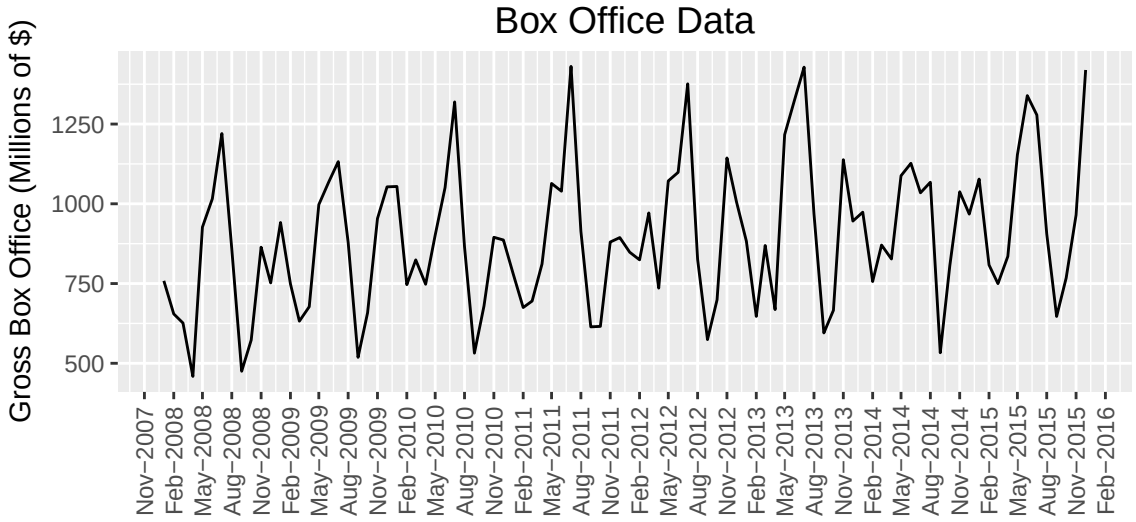

Figure 1: Box Office Plot

The training data were plotted in Figure 1. The trend component isn't obvious but the seasonal component exists and doesn't vary a lot across time. Gross box office peaks every year during summer (May – Aug) and again in winter (Nov – Dec). In addition, the mean of the series is relatively constant. There isn't any significant issue with heteroscedasticity and therefore no Box-Cox transformation is necessary.

3.2 Transformation to Stationary Process

Figure 2 presents the result of an additive decomposition of the training data. Interestingly, the plot shows that the monthly gross box office actually has an upward trend from 2008 to 2015. This result suggests that either box office data cannot explain the decline of the film industry well, or the film industry isn't doing as bad as people think. Besides, the series is nonstationary due to existences of trend and seasonal components, and thus was differenced twice to achieve stationarity.

3.3 ARIMA Model

Potential ARIMA models are selected through checking ACF and PACF plots of the training set (Figure 3), and through ARIMA subset selection (Figure 4). AR(1) can be a good fit because ACF of the series cuts off after lag 1, while PACF decays to 0 through a declined sine wave. Alternatively, ACF of the series decays to 0 through a declined sine wave while PACF cuts off after lag 1, and thus a MA(1) model also fits the data. ARMA(1,1) can be a potential fit as well since both ACF and PACF decay to 0 as time proceeds. In addition, ARMA subset selection suggests fitting a MA(12) model with parameters only at lag 1 and lag 12.

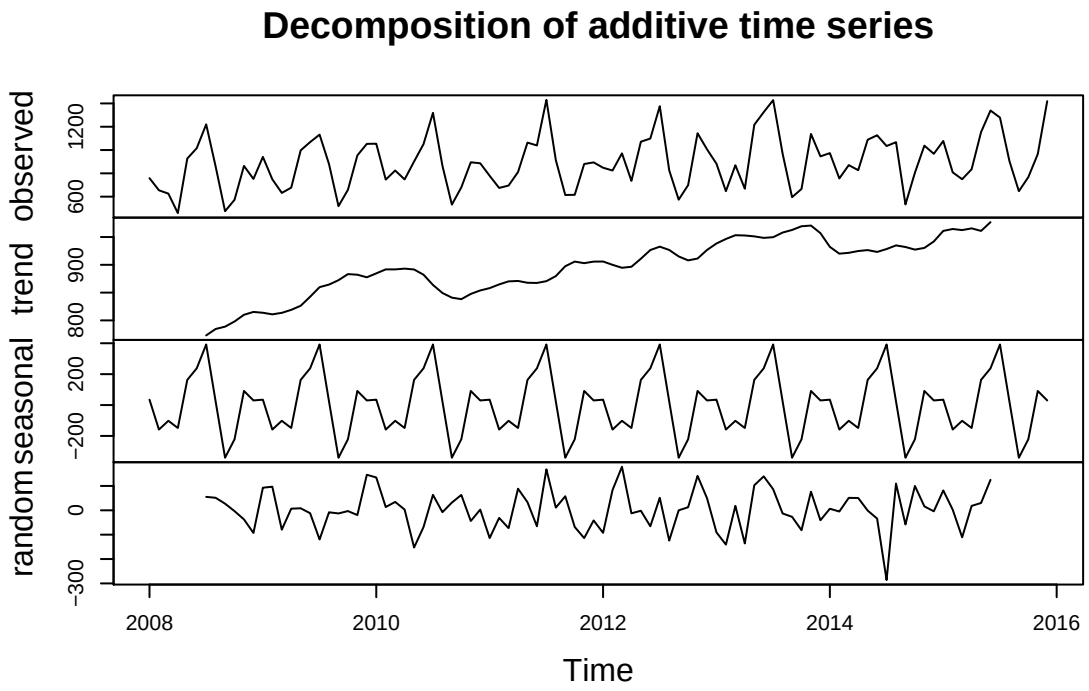

Figure 2: Data Decomposition

A comparison of four potential ARIMA models along with parameter estimations are presented in Table 1. ARIMA(0,2,1) is selected for further analysis since it has the smallest AICc value.

Table 1: Comparison of Fitted ARIMA models

Model	**Full Model**	**AICc**
ARIMA(1,2,0)	$X_t + 0.396X_{t-1} = e_t$	1353.7014446
ARIMA(0,2,1)	$X_t = e_t - 0.999e_{t-1}$	1310.9919323
ARIMA(1,2,1)	$X_t - 0.011X_{t-1} = e_t - 0.999e_{t-1}$	1313.1149988
ARIMA(0,2,12)	$X_t = e_t + 0.367e_{t-12}$	1313.2399042

3.3.1 Diagnostics

According to residual analysis plot (Figure 5), residuals of selected ARIMA model distribute randomly around the x-axis. ACF of residuals mostly lie within the 95% confidence bound. Almost all p-values produced are smaller than 0.05. The normal Q-Q plot for residuals of the model (Figure 6) shows that majority of the points follow the straight line, providing additional evidence that residuals come from approximately normal distribution. The selected model is adequate because its residuals approximately follow a white noise process.

3.4 SARIMA model

Potential SARIMA models are identified through checking ACF and PACF plots of the training set (Figure 3). Since ACF cuts off to zero after the first seasonal lag (lag 12) while PACF at seasonal lags gradually decay to zero, a MA(1) model shall be fitted for the seasonal part. The non-seasonal ARMA(p,q) component can be identified as AR(1), MA(1) or ARMA(1,1) model, and the identification process is exactly the same as selecting orders for ARIMA models outlined in the previous section.

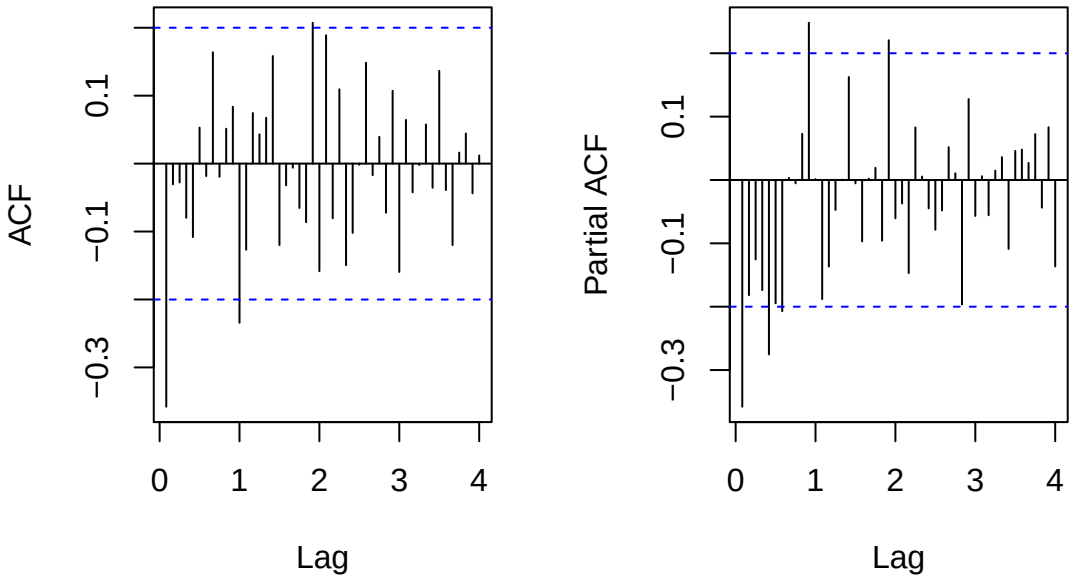

Figure 3: ACF and PACF Plots for Training Data

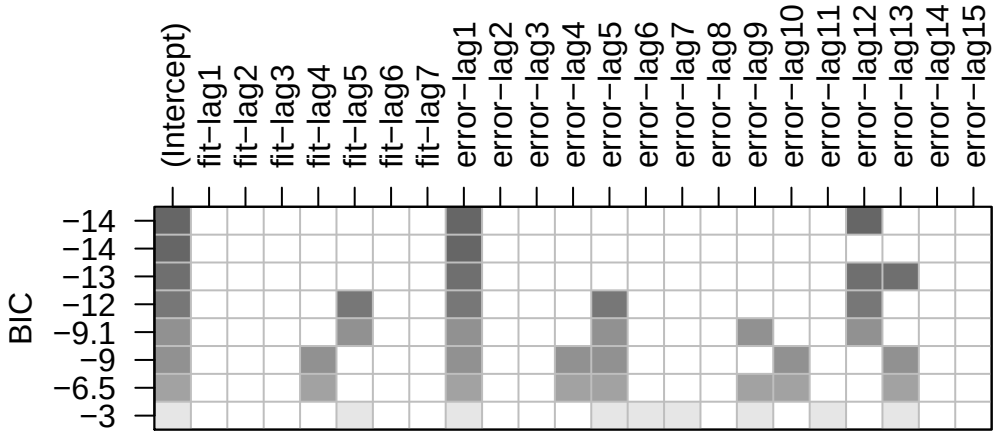

Figure 4: ARIMA subset selection

A comparison of three potential SARIMA models along with estimations of parameters are shown in Table 2. First two models in the table have very close AICc values, which are both smaller than that of the third model, and thus these two models are selected for further analysis.

Table 2: Comparison of Fitted SARIMA Models

Model	**Parameter Estimations**	**AICc**
SARIMA$(0,1,1)\times(0,1,1)_{12}$	$\theta=-0.999$, $\Theta=-0.9998$	1032.6
SARIMA$(1,1,1)\times(0,1,1)_{12}$	$\phi=0.0257$, $\theta=-0.999$, $\Theta=-0.9977$	1034.8
SARIMA$(1,1,0)\times(0,1,1)_{12}$	$\phi=-0.5124$, $\Theta=-0.9980$	1054.3

3.4.1 Diagnostics

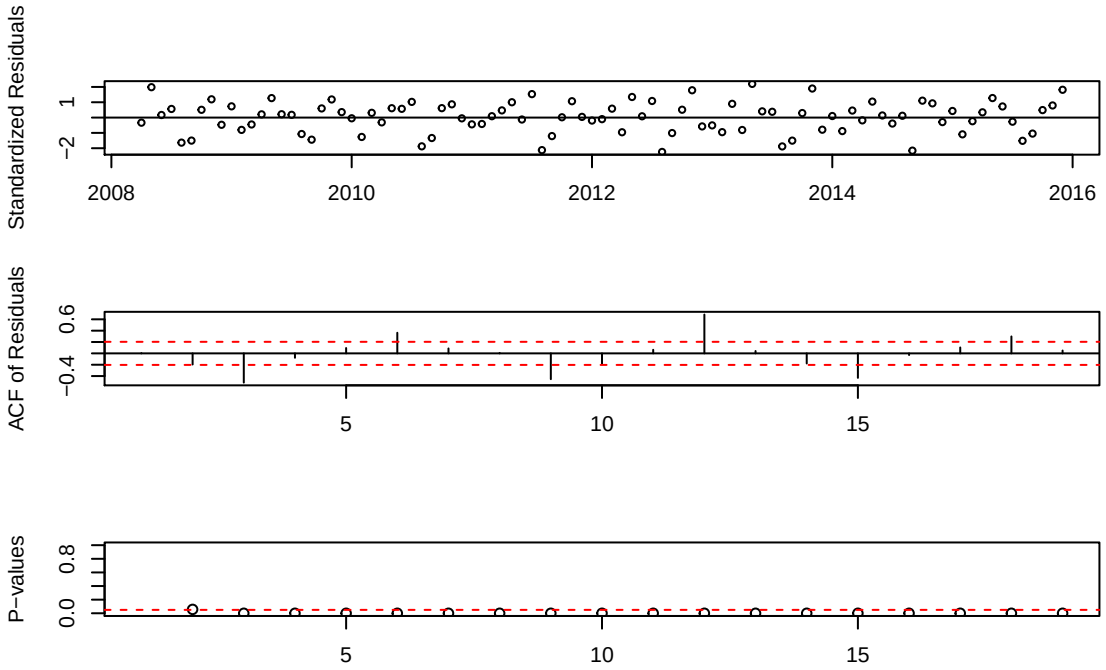

Figure 5: Residual Analysis for ARIMA(0,2,1)

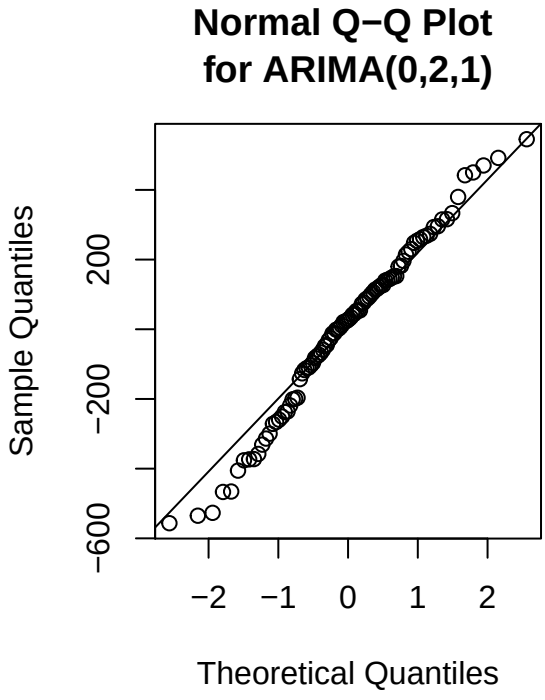

Figure 6: Normal Q-Q Plot of residuals for Fitted ARIMA Model

Figure 7 and 8 are residual analysis plots for SARIMA$(0,1,1)\times(0,1,1)_{12}$ and SARIMA$(1,1,1)\times(0,1,1)_{12}$, respectively. According to standardized residuals plots, residuals of both models distribute randomly around x-axis. ACF of residuals of both models fall within the 95% confidence region. A comparison of normal Q-Q plot for residuals of both models (Figure 9) look very similar to each other, leading to the conclusion that both models are adequate because their residuals are approximate realizations from a white noise process.

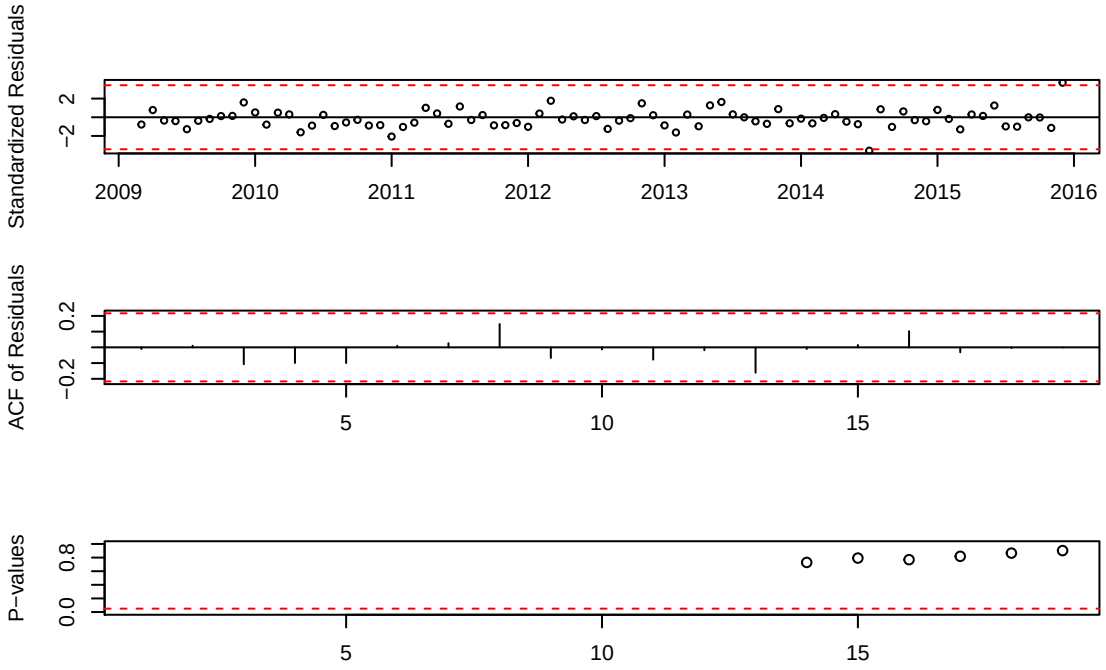

Figure 7: Residual Analysis for SARIMA(0,1,1)(0,1,1)[12]

4 Forecast

As the final step for selecting the best model, potential models' forecast accuracies shall be measured and compared. A forecast for the monthly gross box office from January 2016 to March 2918 was carried out using both Holt-Winters method and ARMA models selected from previous sections. Forecast errors are measured by the difference between observed values in the test set and forecast values.

Criteria for forecast accuracy such as RMSE, MAPE and MAE were calculated through forecast errors and the result of the comparison is presented in Table 3. It can be inferred that SARIMA forecasts perform significantly better than both ARIMA and Holt-Winters forecasts because of smaller values of RMSE, MAPE and MAE. Besides, SARIMA_2 has slightly more advantages than SARIMA_1. Thus, SARIMA$(1,1,1)\times(0,1,1)_{12}$ is chosen as the final model. Figure 10 is a forecast for monthly gross box office from April 2018 to March 2020 with the finalized model. The light gray region is the 95% prediction interval while the dark gray area acts as the 80% prediction interval. It is forecasted that monthly gross box office will have a flattened trend rather than an upward trend possessed by the training set. This predicted change in behavior of box office may suggest the decline of the US film industry.

Table 3: Comparison of Forecast Accuracy

Criteria	**ARIMA(0,2,1)**	**SARIMA_1**	**SARIMA_2**	**Holt-Winters**
RMSE	543.3922728	157.9594917	157.7788706	236.6948658

Criteria	ARIMA(0,2,1)	SARIMA_1	SARIMA_2	Holt-Winters
MAPE	55.7313478	11.5721819	11.5442261	21.5616005
MAE	489.7107046	116.6550199	116.2925189	198.4102346

(*Note:* SARIMA_1: SARIMA$(0,1,1) \times (0,1,1)_{12}$; SARIMA_2: SARIMA$(1,1,1) \times (0,1,1)_{12}$)

5 Discussion

Although the trend of box offcie is predicted to decline after April 2018, the fitted model is incompatible with the argument of decline of US film industry in recent years due to the upward trend of training data (Figure 3). The predictive power of box office to the movie industry should be further confirmed. One possible reason that the series analyzed in this paper has an upward trend is that the effect of rising ticket price wasn't taken into account. Box office data might behave very differently, such as displaying a downward trend, if it is adjusted for ticket prices. Therefore, it's recommended for future studies to modify the box office data based on price effect before any form of analysis.

In addition, Plaugic (2018) reports that the US movie theater attendance experienced its 25-year low in 2017. However, as shown in Figure 12, gross box office in 2017 wasn't affected very much by the shock as it remains at roughly the same level as gross box office in 2016. It is thus speculated that the rising ticket price serves as a cushion for low attendance rate, stabilizing 2017 box office. Comparing forecast of monthly box office from January 2016 to March 2018, which are plotted with the blue line, and actual observed values, which are plotted with the red dash line (Figure 11), it can be seen that most actual values in 2017 fall within 95% prediction interval. This result suggests that predictions made for 2017, though not as accurate as they are expected to be, are generally acceptable. Yet, if the low attendance rate were to continue with rising ticket price, the predictive power of the current model would be serious undermined as it doesn't capture these two effects. As a result, future studies are strongly encouraged to check if the low attendance rate in 2017 is a permanent or transitory shock to the show business, and if the shock is permanent, data should be modified accordingly to reflect ticket sales.

6 Conclusion

Many recent articles express concerns for declining US film industry due to threats from rising ticket prices and increasing competition from streaming services. It's thus important to forecast the future performance of the industry so that related sectors can adjust for changes accordingly, and the study strives to develop a model that would produce reliable forecasts with recent data. The paper intends to analyze monthly box office data from January 2008 to March 2018 using time series approach, and both ARIMA models and SARIMA models were fitted to the data and selected using selection criteria. Selected models were also verified for adequacy through residual analysis. In addition, forecasts carried out by selected ARMA models were compared with forecasts by Holt-Winters method, and the one with highest forecast accuracy was identifies as the best fit. The resulting model is SARIMA$(1,1,1) \times (0,1,1)_{12}$ with parameters $\phi = 0.0257, \theta = -0.999, \Theta = -0.9977$. However, the stability of the model is in uncertainty due to movie theater attendance rate shock in 2017 and risking ticket prices. Future analysis is thus recommended to adjust for ticket sales and price effect.

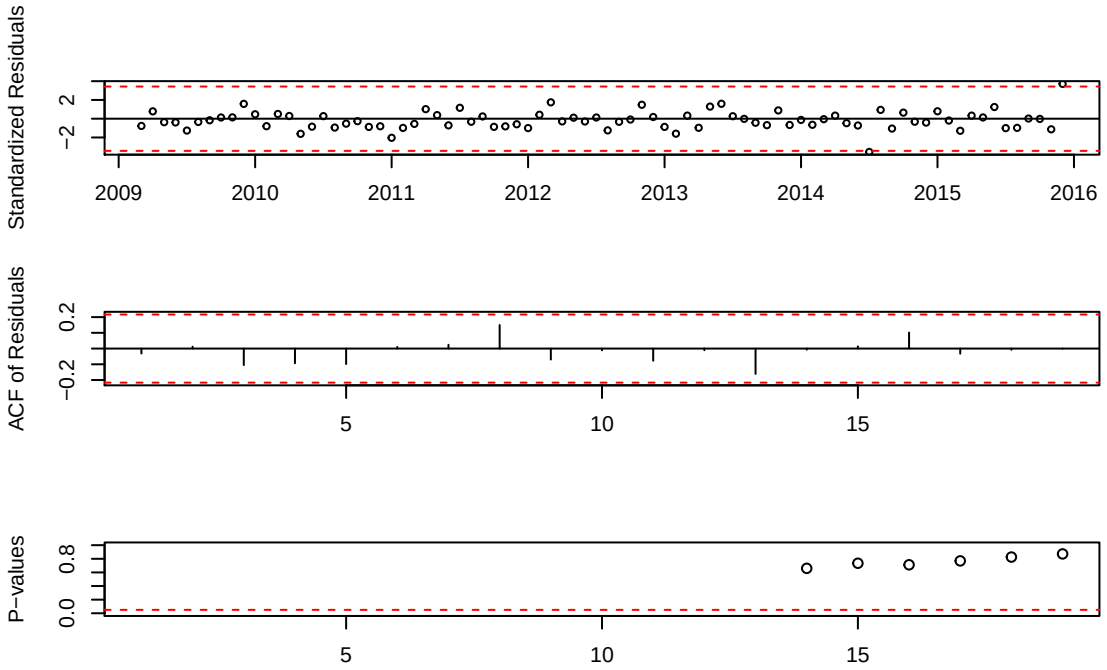

Figure 8: Residual Analysis for SARIMA(1,1,1)(0,1,1)[12]

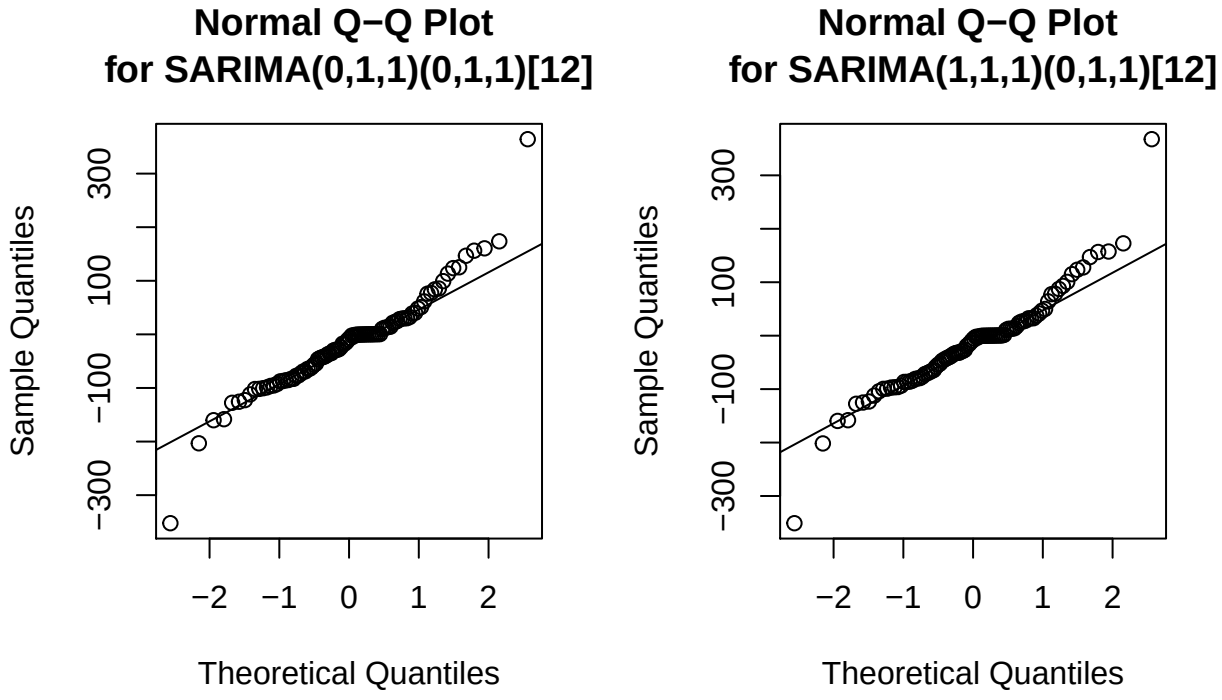

Figure 9: Normal Q-Q Plots of Residuals for Two Fitted SARIMA Models

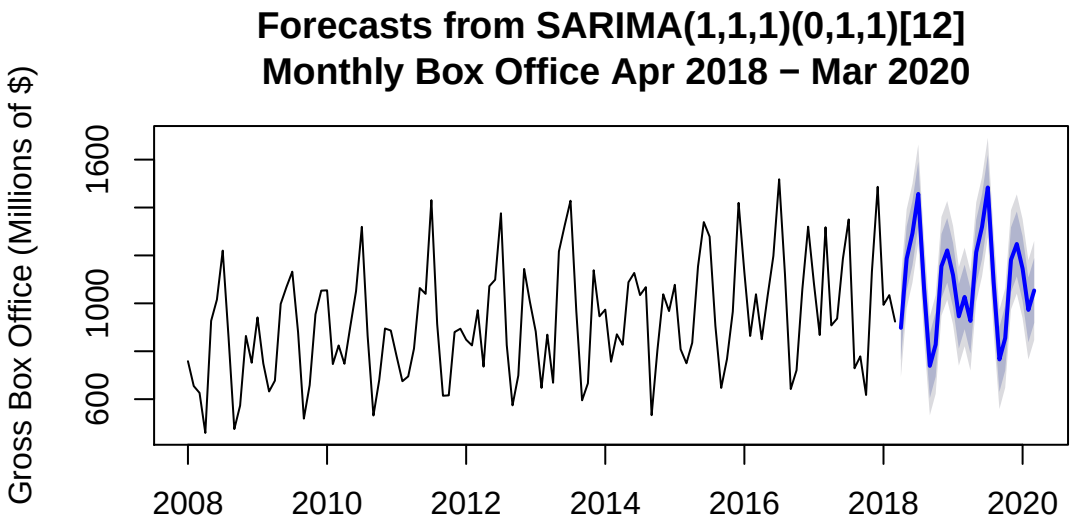

Figure 10: Monthly Box Office Forecasts for April 2018 - March 2020

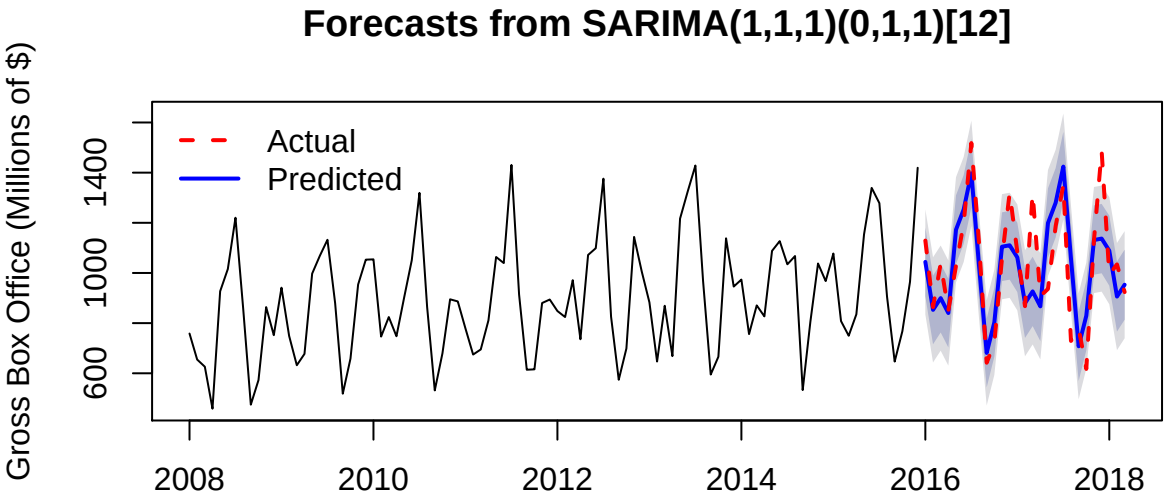

Figure 11: Forecast Accuracy - Comparison between Predicted Values and Observed Values

7 Reference

Bilton, Nick. 2017. "Why Hollywood as We Know It Is Already over." *Vanity Fair*. https://www.vanityfair.com/news/2017/01/why-hollywood-as-we-know-it-is-already-over.

Plaugic, Lizzie. 2018. "Domestic Movie Theater Attendance Hit a 25-Year Low in 2017." *The Verge*. https://www.theverge.com/2018/1/3/16844662/movie-theater-attendance-2017-low-netflix-streaming.

Statista. 2018. "Film and Movie Industry - Statistics & Facts." https://www.statista.com/topics/964/film/.