**Forecasting the number of large-scale strikes in the USA during 1947 – 2020**

**Abstract**

The main aim of this study was to evaluate the long-term movement of the number of large-scale strikes occurring annually in the USA and to forecast the number of strikes for the years 2021 and 2022. The publicly available data on strikes during the period 1947 to 2020 were collected from the bureau of labor statistics website.

Box-Jenkin’s methodology was employed to determine the appropriate form of the model to explain the behavior of large-scale strikes. The results of the ARIMA modeling initially indicated log-transformed, first-order differenced with a mix of auto-regressive and moving average component model, ARIMA (2, 1, 2). The comparison of alternate ARIMA models indicated that ARIMA (2, 1,1) performed better and was considered the final model for forecasting. The analysis indicated that there was no deterministic trend in the number of strikes. The forecasted number of strikes for 2021 and 2022 was small, less than 3.

**Introduction**

The impact of work disruptions due to strikes across industries has attracted extensive attention from researchers and policy decision-makers. Large-scale strikes involving a large number of workers have a multidimensional impact on economic, political, social, and community-related matters. Large-scale strikes not only harm the economic output but can also potentially result in social and political changes.

The main aim of this study is to evaluate the long-term patterns in the number of large-scale strikes or work stoppages involving thousand or more workers in the USA and to develop a forecast for 2021 and 2022. The time-series data on strikes was collected from the bureau of labor statistics website, <https://www.bls.gov/web/wkstp/annual-listing.htm>. The data contains time series data of the number of work stoppages from 1947 to 2020.

The time series analysis was performed using the univariate time series analysis approach, specifically using Box-Jenkin’s methodology. The Auto-Regressive Integrated Moving Average (ARIMA) was employed to identify the appropriate form of the model. The form of the ARIMA model is determined by order of auto-regression (p), order of moving average (q), and the order of differencing that makes the times series stationary (d). The initial time series plot was assessed for the need for any transformation to achieve variance stabilization. The autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to determine the values of d, p, and q for the initial candidate model. The model adequacy analysis was performed using the Ljung-Box test for white-noise residuals. Alternative candidate models were fitted by changing the values of p and q and model performance was assessed using AIC measure for model parsimony. The ARIMA model with the least value of AIC was considered the parsimonious model and used for the final forecasting of strikes for the years 2021 and 2022. The ARIMA modeling and the associated analysis were performed in R, version 4.1.2.

**Results**

The large-scale strikes from 1947 to 2020 ranged between a minimum of 5 in 2009 and a maximum of 470 in 1952 with a mean = 156.11 strikes (SD = 145.72) and a median of 75. Figure 1 is the time series plot of strikes. This plot indicates non-constant variance in strikes across the period from 1947 – 1980. Specifically, the time series for 1947 – 1980 indicated higher volatility compared with post-1980’s data. Therefore, the strikes time series were log-transformed to achieve variance stabilization. Figure 2 is the time series plot of the data of the log-transformed strike. This plot indicates a stable variance and general decreasing trend in the strikes. The log-transformed series was first differenced to check for stationarity. The plot of the first differenced log-transformed strikes in figure 3 indicates stationary series.

Figures 4 and 5 respectively are the ACF and PACF plots of the first order differenced log-transformed strikes data. The ACF plot indicates significant autocorrelations at lags 1 and 2. The PACF plot too indicates significant partial autocorrelations at lags 1 and 2. These results indicate a mixture of autoregressive (AR) and moving-average (MA) components in the initial ARIMA model with first-order integration. Therefore, the initial candidate ARIMA model for strikes during 1947 – 2020 was set as ARIMA (2, 1, 2). The estimated AR coefficients in the ARIMA (2, 1, 2) model was ϕ1 = -0.38 (SE = .457), ϕ2 = -0.22 (SE =.25) and the MA coefficients were θ1 = 0.064 (SE = .45) and θ1 = -0.117 (SE = .28).

Three candidate models were fitted and compared with the initial form of the model ARIMA (2, 1, 2). These candidate models were – (i) ARIMA (1, 1, 1), (ii) ARIMA (2, 1, 1) and (iii) ARIMA (1, 1, 2).

Table 1 presents the model comparison statistics of the initial ARIMA model and the alternate candidate models. ARIMA (2, 1, 1) model reports the least value of AIC and RMSE. Therefore, based on AIC and RMSE measures, ARIMA (2, 1, 1) was found to be the most parsimonious model and it was considered as the final model and used for forecasting the number of strikes for the years 2021 and 2022.

**Table 1.** *Comparison of alternate ARIMA models*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Form | AIC | RMSE | MAPE |
| Model 1 (initial) | ARIMA (2, 1,2) | 61.15 | .341 | 7.789% |
| Model 2 | ARIMA (1, 1, 1) | 59.75 | .347 | 7.941% |
| Model 3 | ARIMA (2, 1, 1) | 59.29 | .341 | 7.821% |
| Model 4 | ARIMA (1, 1, 2) | 59.87 | .342 | 7.824% |

Figure 6 is a time series plot of the residuals from the final ARIMA (2, 1, 1) model. The plot indicates a resemblance to a white-noise process. The histogram of the residuals from ARIMA (2, 1, 1) in figure 7 indicated symmetric distribution. The results of the Ljung-Box test indicated evidence for white-noise residuals (χ2 (12) = 12.04, p = .443). Therefore, the final model ARIMA (2, 1, 1) passes the model adequacy test. The final model was chosen for the forecasting of the number of strikes for the years 2021 and 2022 is ARIMA (2,1, 1). The estimated AR coefficients in the final model was ϕ1 = -0.239, ϕ2 = -0.292 and the MA coefficient was θ1 = -0.079. This model in algebraic form using the backshift operator is -

(1 – ϕ1B – ϕ2B2) (1-B) Yt = (1 – θ1B) εt where BpYt  = Yt-p, the pth lag of Yt and (1-B)d is the dth ordered difference of Yt.

The fitted ARIMA model is: (1 – (-.239) B – (-.292) B2) (1-B) Yt = (1 – (-0.079) B) εt.

On simplification, the fitted ARIMA (2, 1, 1) model is as follows.

Yt - 0.761 Yt-1 + 0.053 Yt-2 – 0.292 Yt-3 = εt + 0.079 εt-1

Figure 8 is a plot of the actual and forecasted number of strokes during the period 1947 – 2020. Figure 9 is the plot of the forecast for the years 2021 and 2022. Using the ARIMA (2, 1, 1) model, the forecast of the number of strikes for the year 2021 was 2.346 (95% CI: 1.674 – 3.019) and for the year 2022, 2.614 (95% CI: 1.800 – 3.429).

**Conclusions**

The main aim of this study was to evaluate the long-term movement of the number of large-scale strikes occurring annually in the USA and to forecast the number of strikes for the years 2021 and 2022. The publicly available data on the number of strikes during the period 1947 to 2020 was collected from the bureau of labor statistics website.

Box-Jenkin’s methodology was employed to determine the appropriate form of the model to explain the behavior of large-scale strikes. The results of the ARIMA modeling initially indicated log-transformed, first-order differenced with a mix of auto-regressive and moving average component model, ARIMA (2, 1, 2). The comparison of alternate ARIMA models indicated that ARIMA (2, 1,1) performed better and was considered the final model for forecasting. It was concluded that indicated that there was no deterministic trend the number of strikes and the forecasted number of strikes for 2021 and 2022 were less than 3.

Chart, histogram

Description automatically generated

Figure 1. Time series plot of the number of strikes during 1947 – 2020

Chart, histogram

Description automatically generated

Figure 2. Time series plot of the logged number of strikes during 1947 – 2020

Graphical user interface, chart

Description automatically generated

Figure 3. Time series plot of the first differenced log strikes during 1947-1960

Chart

Description automatically generated

Figure 4. ACF curve of the first differenced log number of strikes during 1947 -2020

Chart, box and whisker chart

Description automatically generated

Figure 5. PACF curve of the first differenced log number of strikes during 1947 – 2020

Chart

Description automatically generated

Figure 6. Time series plot of the residuals from ARIMA(2, 1, 1) model

Chart, histogram

Description automatically generated

Figure 7. Histogram of residuals from ARIMA(2, 1,1) model

Chart, histogram

Description automatically generated

Figure 8. Time series plot of the observed and forecasted number of strikes during 1947 – 2020

Chart, histogram

Description automatically generated

Figure 9. Forecasted log strikes for 2021 and 2022

**Appendix**

**R Program**

# clear the memory space

rm(list=ls())

#read the data file

strikes0<- read.csv(file.choose(), header=T)

str(strikes0)

library(psych)

describe(strikes0$Strikes)

#define the time series settings

strikes<- ts(strikes0$Strikes, frequency=1, start=1947)

str(strikes)

# plot the time series plot

plot.ts(strikes, ylab="Number of strikes during 1947 - 2020")

# log of strikes and its plot

logstrikes<- log(strikes)

plot.ts(logstrikes, ylab="Logarithm of strikes during 1947 - 2020")

library(forecast)

# differenccing the time series data and checking for stationarity

logstrikes\_diff<- diff(logstrikes, differences=1)

plot.ts(logstrikes\_diff)

plot.ts(logstrikes\_diff, ylab="First differenced log strikes during 1947 - 2020")

# ACF and PACF of the first differenced logarithm of strikes

acf(logstrikes\_diff, main="", ylab="Auto correlation of first differenced log strikes")

pacf(logstrikes\_diff, main="", ylab="Partial auto correlation of first differenced log strikes")

# fitting the intial ARIMA candidate model - ARIMA (2,1,2)

arima0<- arima(logstrikes, order=c(2, 1, 2))

summary(arima0)

# fitting and checking alternate candidate models

arima1<- arima(logstrikes, order=c(1, 1, 1))

summary(arima1)

arima2<- arima(logstrikes, order=c(2, 1, 1))

summary(arima2)

arima3<- arima(logstrikes, order=c(1, 1, 2))

summary(arima3)

model=c("arima212", "arima111", "arima211", "arima112")

aic=c(arima0$aic, arima1$aic, arima2$aic, arima3$aic)

modelcompare<- cbind(model, aic)

modelcompare

arima2\_forecast<- forecast(arima2, h=2, level=95)

plot(arima2\_forecast, ylab="Forecast of the log strikes for 2021 and 2022", main="")

plot.ts(arima2$residuals, ylab="Residuals of ARIMA(2, 1, 1) model")

Box.test(arima2$residuals, lag=12, type="Ljung-Box")

hist(arima2$residuals, xlab=" Residuals of ARIMA(2, 1, 1)", col="gold", main="")

plot(exp(arima2\_forecast$x), ylab="Number of strikes", col="blue")

lines(exp(arima2\_forecast$fitted), col="red")

legend("topright", legend=c("Observed", "Forecast"), col=c("blue", "red"), lty=1, cex=0.8)