Advanced Time Series Analysis and Forecasting with SARIMA

Time series model:

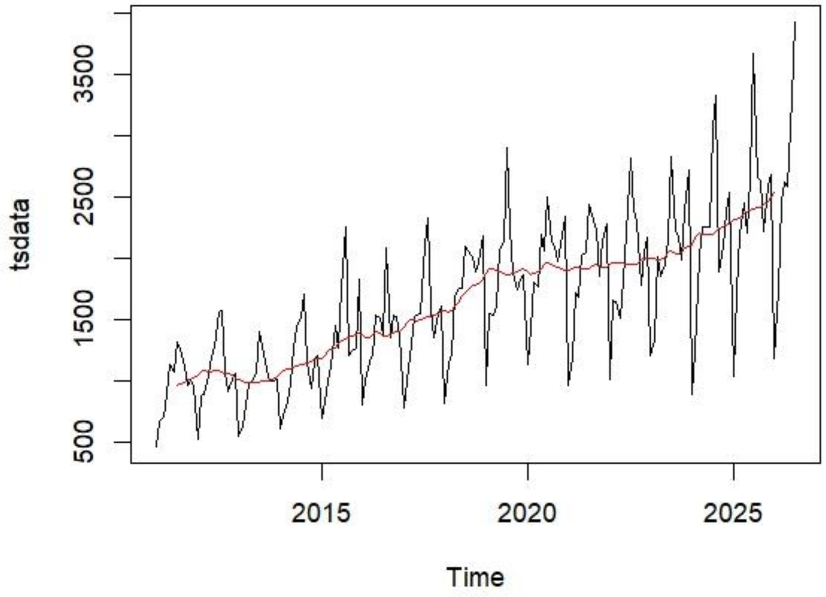


Fig.1

A time series from around 2010 to 2025 is shown in the plot (Fig.1). The red line represents the moving average, and the black line represents actual fluctuating data.

Checking the stationarity of my series:

ACF plot:

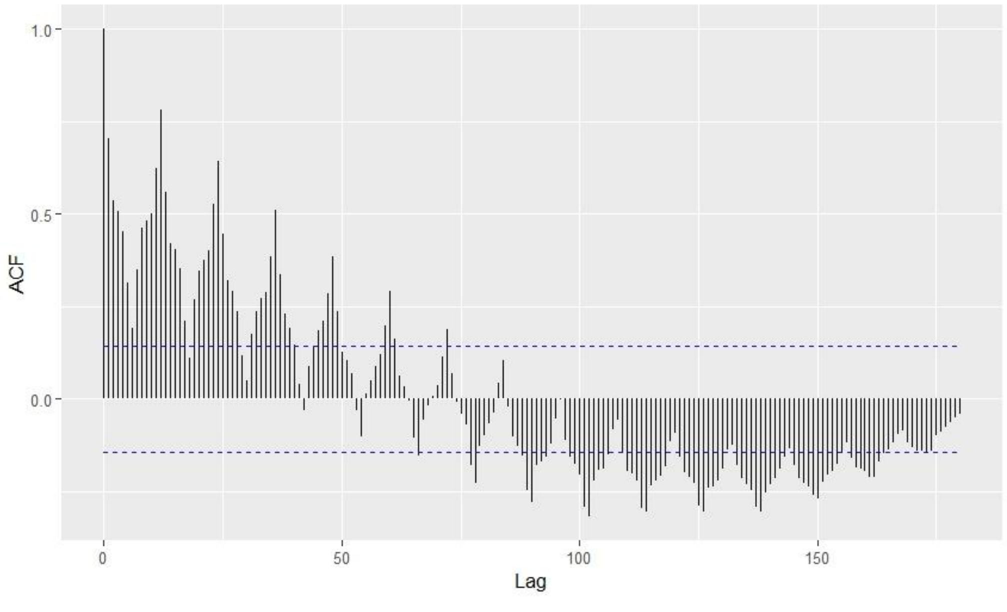


Fig.2

The autocorrelations in the ACF plot (Fig.2) are slowly decreasing with the increase in lags, which is a common feature of a non-stationary series.

PACF plot:

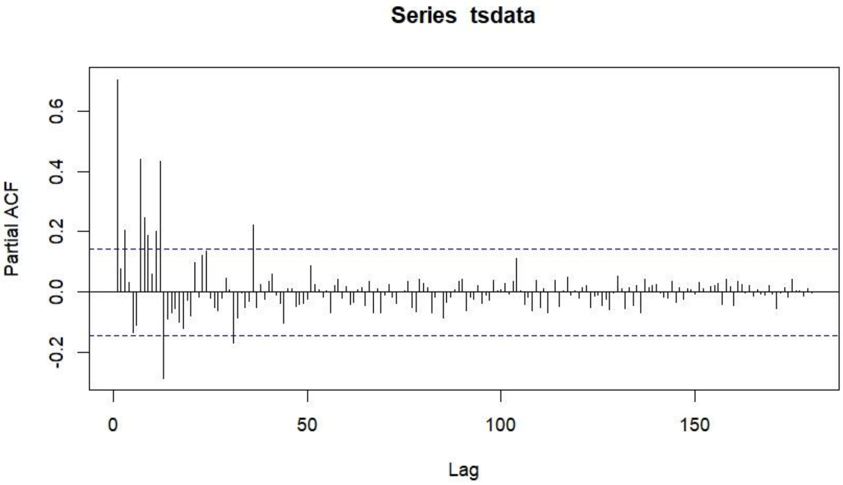


Fig.3

The extended decay in the ACF is more suggestive of non-stationarity than anything else. The PACF (Fig.3), shows a sharp cut-off after a few lags, points to an underlying autoregressive structure.

Differencing and Transformation:

Logarithmic transformation:

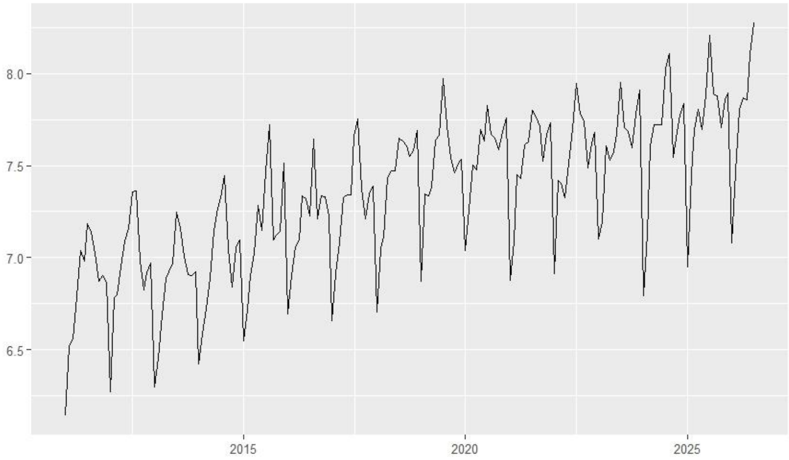
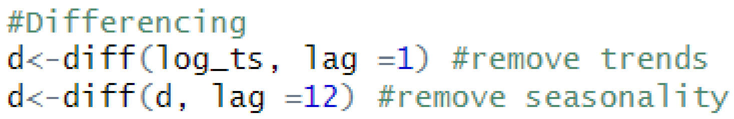


Fig.4

By applying the logarithmic transformation, the variance should be stabilized across the time series. In the transformed series (Fig.4), we can see a reduction in the variability and more symmetric series compared to the original time series shown in Fig.1.

Differencing:



```
#Differencing
d<-diff(log_ts, lag =1) #remove trends
d<-diff(d, lag =12) #remove seasonality
```

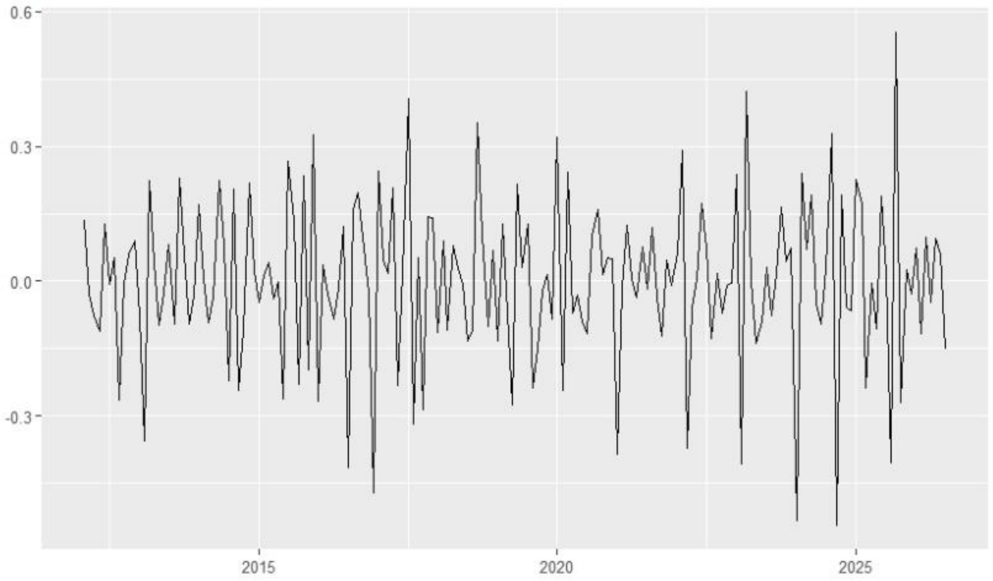
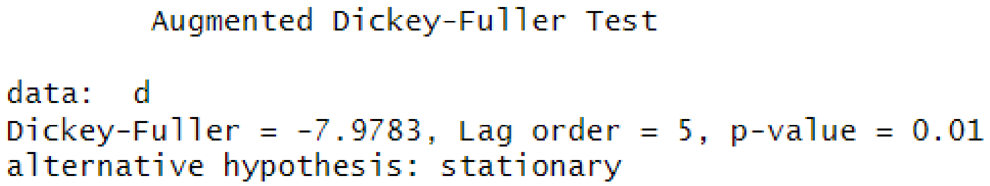


Fig.5

The differencing is applied twice to the logarithmic transformed series to achieve the above stationary plot (Fig.5). The twice-differenced series exhibits no trend or seasonality, but instead fluctuates around a mean of zero.

- First differencing: diff(log_ts, lag = 1) applies the first differencing on the logarithmically transformed time series (log_ts). This is used to remove the trends in the series.
- Second differencing: diff(d, lag = 12) applies a second differencing with a lag of 12 to the first differencing result. To get rid of seasonal patterns.

Augmented Dickey-Fuller (ADF) test:



```
            Augmented Dickey-Fuller Test

data:  d
Dickey-Fuller = -7.9783, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Fig.6

Since the p-value is significantly low and the ADF statistic is negative, we can confidently reject the null hypothesis. This test result suggests that my time series is stationary after the applied transformation and differencing.

ACF after differencing:

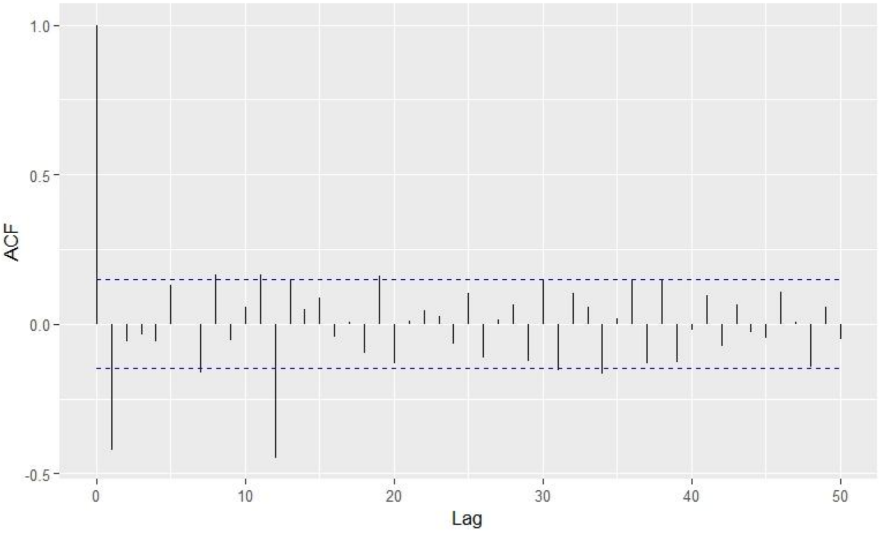


Fig.7

After the differencing, the ACF plot indicates that after the first few lags, the autocorrelations rapidly fall below the threshold (within blue dashed lines). This suggests that after transforming and differencing, there is very little autocorrelation left.

PACF after differencing:

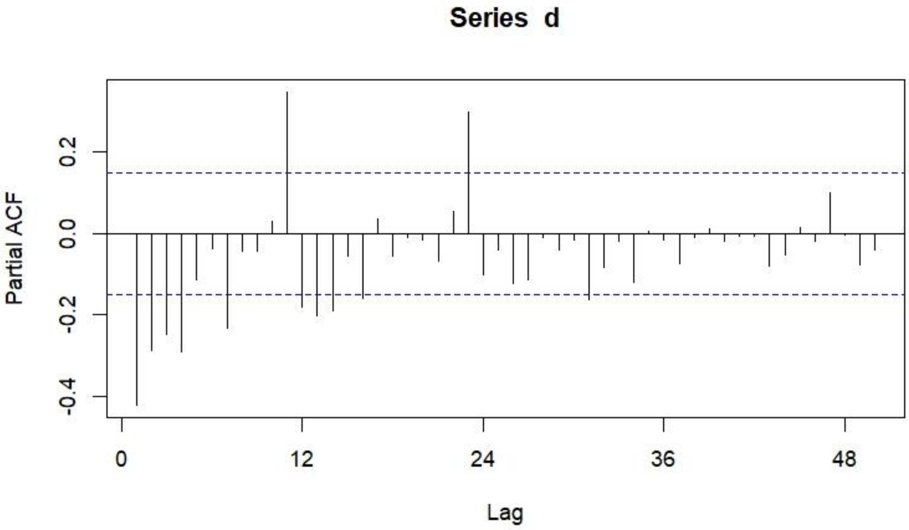


Fig.8

In the PACF plot. Following a significant spike at the first lag, the subsequent spikes mostly stay inside the significance bounds indicated by the blue dashed lines.

Determining the order of SARIMA:

Identifying the order of non-seasonal components AR (p), MA (q), and Differencing(d)

- In PACF, the lags after which PACF cuts off significantly suggest the order of ‘p’, which could be 4.
- In ACF, the lag at which ACF cuts off sharply, tells us the order of the component ‘q’ that is 1. ACF does not exhibit any clearly noticeable spikes other than those close to the initial lags.
- The ‘d’ will be 1 as the series needs to be differenced for the stationarity.

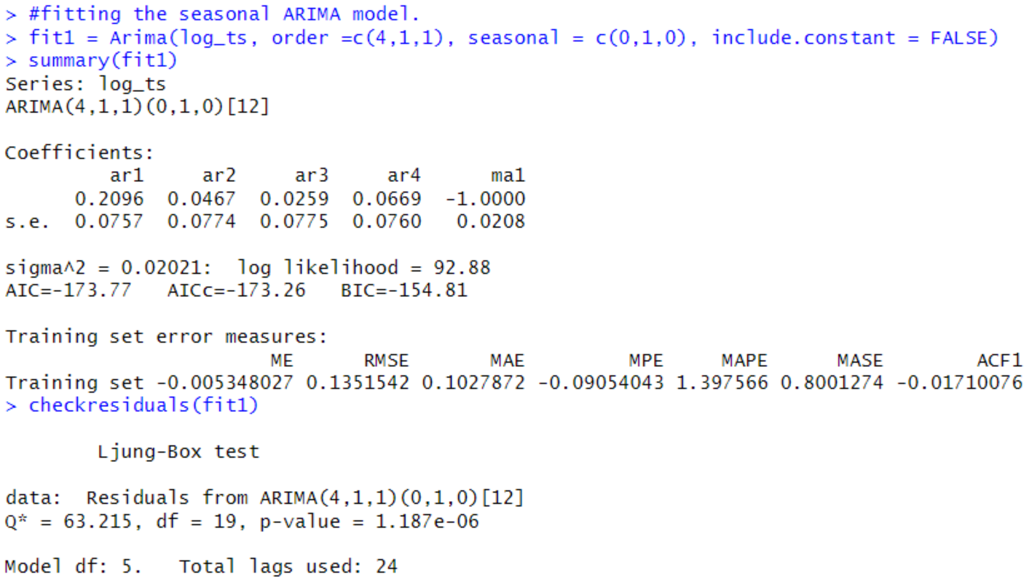
  The order of non-seasonal components (p,d,q) is (4,1,1) from the acf and pacf plots.

Identifying the order of seasonal components AR (P), MA (Q), and Differencing(D).

- Neither ACF nor PACF plots exhibit a distinct seasonal pattern at multiples of 12 months, hence, P = 0 and Q = 0 can be considered.
- The ‘D’ will be 1 as the series needs to be differenced for the stationarity.

  The order of seasonal components (P,D,Q) is (0,1,0) from the acf and pacf plots.

Fitting the seasonal ARIMA model based on the above determined order:



```
> #fitting the seasonal ARIMA model.
> fit1 = Arima(log_ts, order =c(4,1,1), seasonal = c(0,1,0), include.constant = FALSE)
> summary(fit1)
Series: log_ts
ARIMA(4,1,1)(0,1,0)[12]

Coefficients:
         ar1     ar2     ar3     ar4      ma1
      0.2096  0.0467  0.0259  0.0669  -1.0000
s.e.  0.0757  0.0774  0.0775  0.0760   0.0208

sigma^2 = 0.02021:  log likelihood = 92.88
AIC=-173.77   AICc=-173.26   BIC=-154.81

Training set error measures:
                       ME      RMSE       MAE         MPE     MAPE      MASE        ACF1
Training set -0.005348027 0.1351542 0.1027872 -0.09054043 1.397566 0.8001274 -0.01710076
> checkresiduals(fit1)

        Ljung-Box test

data:  Residuals from ARIMA(4,1,1)(0,1,0)[12]
Q* = 63.215, df = 19, p-value = 1.187e-06

Model df: 5.   Total lags used: 24
```

Fig.9

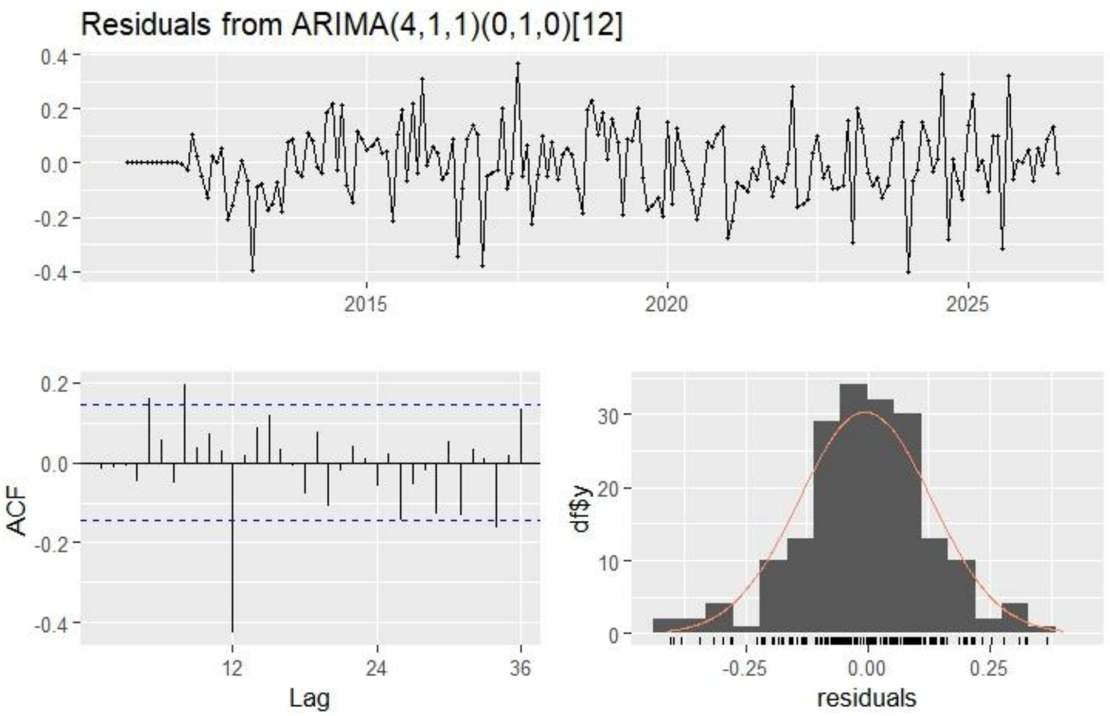
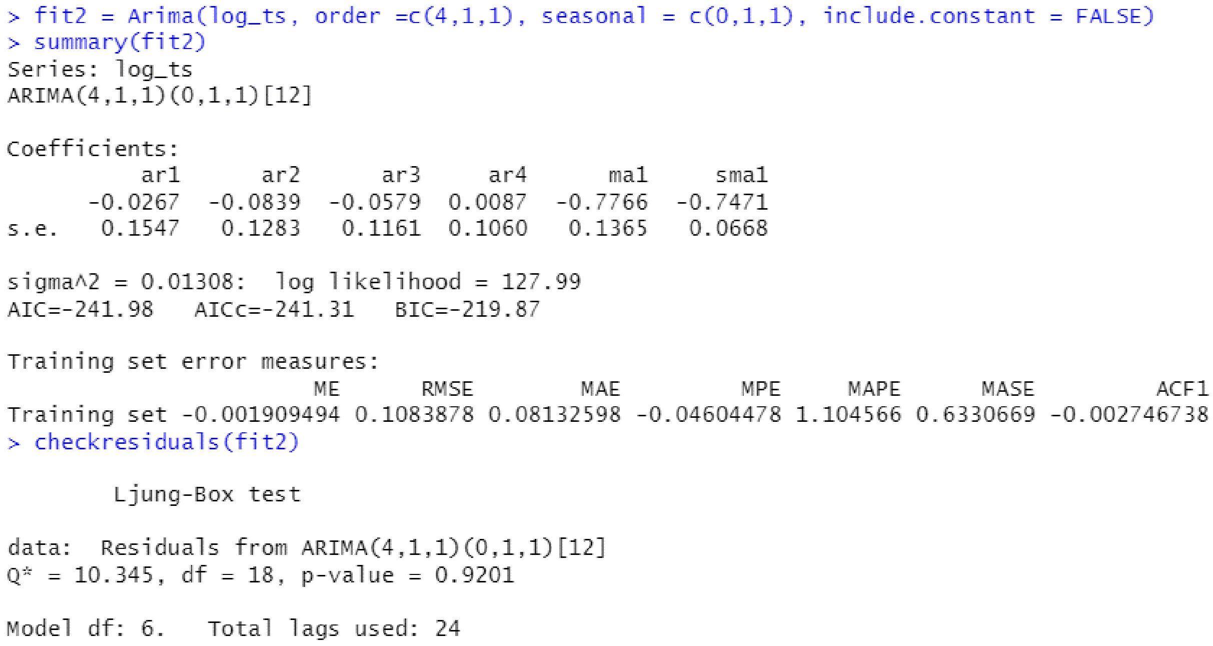


Fig.10

First Possible variation (fit 2):

With order (4,1,1) (0,1,1)

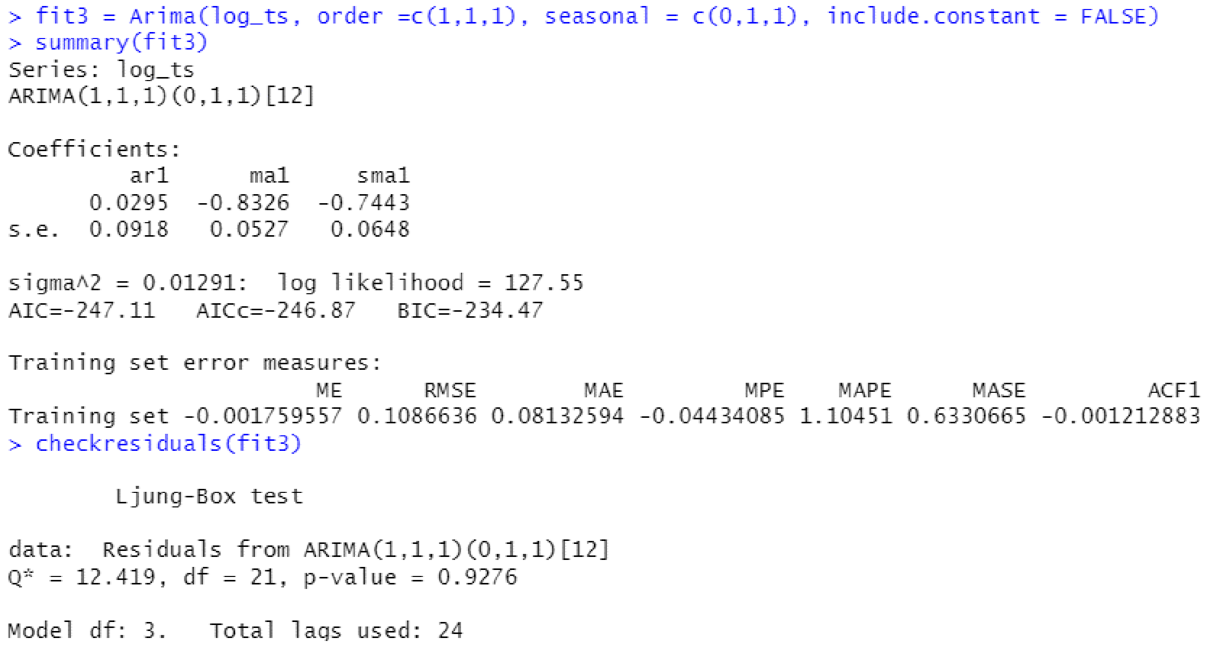


```
> fit2 = Arima(log_ts, order =c(4,1,1), seasonal = c(0,1,1), include.constant = FALSE)
> summary(fit2)
Series: log_ts
ARIMA(4,1,1)(0,1,1)[12]

Coefficients:
          ar1      ar2      ar3     ar4      ma1     sma1
      -0.0267  -0.0839  -0.0579  0.0087  -0.7766  -0.7471
s.e.   0.1547   0.1283   0.1161  0.1060   0.1365   0.0668

sigma^2 = 0.01308:  log likelihood = 127.99
AIC=-241.98   AICc=-241.31   BIC=-219.87

Training set error measures:
                       ME      RMSE        MAE         MPE     MAPE      MASE         ACF1
Training set -0.001909494 0.1083878 0.08132598 -0.04604478 1.104566 0.6330669 -0.002746738
> checkresiduals(fit2)

        Ljung-Box test

data:  Residuals from ARIMA(4,1,1)(0,1,1)[12]
Q* = 10.345, df = 18, p-value = 0.9201

Model df: 6.   Total lags used: 24
```

Fig.11

Second possible variation (fit3):

With order (1,1,1) (0,1,1)



```
> fit3 = Arima(log_ts, order =c(1,1,1), seasonal = c(0,1,1), include.constant = FALSE)
> summary(fit3)
Series: log_ts
ARIMA(1,1,1)(0,1,1)[12]

Coefficients:
         ar1      ma1     sma1
      0.0295  -0.8326  -0.7443
s.e.  0.0918   0.0527   0.0648

sigma^2 = 0.01291:  log likelihood = 127.55
AIC=-247.11   AICc=-246.87   BIC=-234.47

Training set error measures:
                       ME      RMSE        MAE         MPE    MAPE      MASE          ACF1
Training set -0.001759557 0.1086636 0.08132594 -0.04434085 1.10451 0.6330665 -0.001212883
> checkresiduals(fit3)

        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)(0,1,1)[12]
Q* = 12.419, df = 21, p-value = 0.9276

Model df: 3.   Total lags used: 24
```

Fig.12

Determining the best fit after the comparison of all the 3 variations:

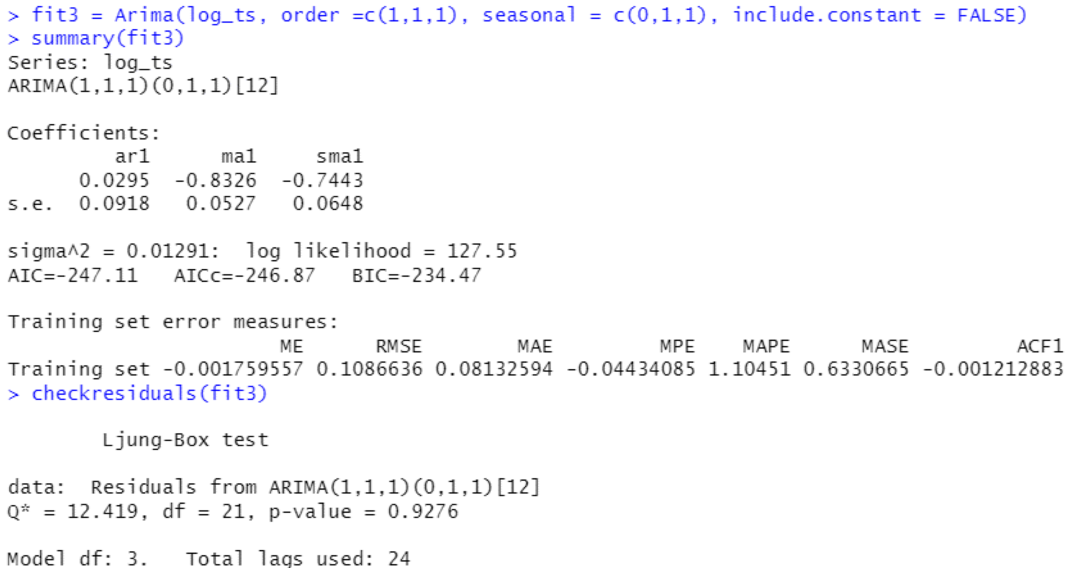
Log-likelihood: Fit2 and Fit3 exhibit significantly higher log-likelihood values in comparison to Fit1, suggesting that they provide a closer fit to the actual data points. Fit2 marginally surpasses Fit3.

AIC and BIC: Lower values are preferable because they show a more successful balance between model complexity and goodness of fit. Fir 3 has lower AIC and BIC makes it a better fit.

p-value: Both Fit2 and Fit3 show no significant autocorrelations. But the lower AIC and the BIC values of the fit3 make it more preferable.

Therefore, fit3 is the best fit.

The Final selected model: ARIMA(1,1,1)(0,1,1)



```
> fit3 = Arima(log_ts, order =c(1,1,1), seasonal = c(0,1,1), include.constant = FALSE)
> summary(fit3)
Series: log_ts
ARIMA(1,1,1)(0,1,1)[12]

Coefficients:
         ar1      ma1     sma1
      0.0295  -0.8326  -0.7443
s.e.  0.0918   0.0527   0.0648

sigma^2 = 0.01291:  log likelihood = 127.55
AIC=-247.11   AICc=-246.87   BIC=-234.47

Training set error measures:
                       ME      RMSE        MAE         MPE    MAPE      MASE         ACF1
Training set -0.001759557 0.1086636 0.08132594 -0.04434085 1.10451 0.6330665 -0.001212883
> checkresiduals(fit3)

        Ljung-Box test

data:  Residuals from ARIMA(1,1,1)(0,1,1)[12]
Q* = 12.419, df = 21, p-value = 0.9276

Model df: 3.   Total lags used: 24
```

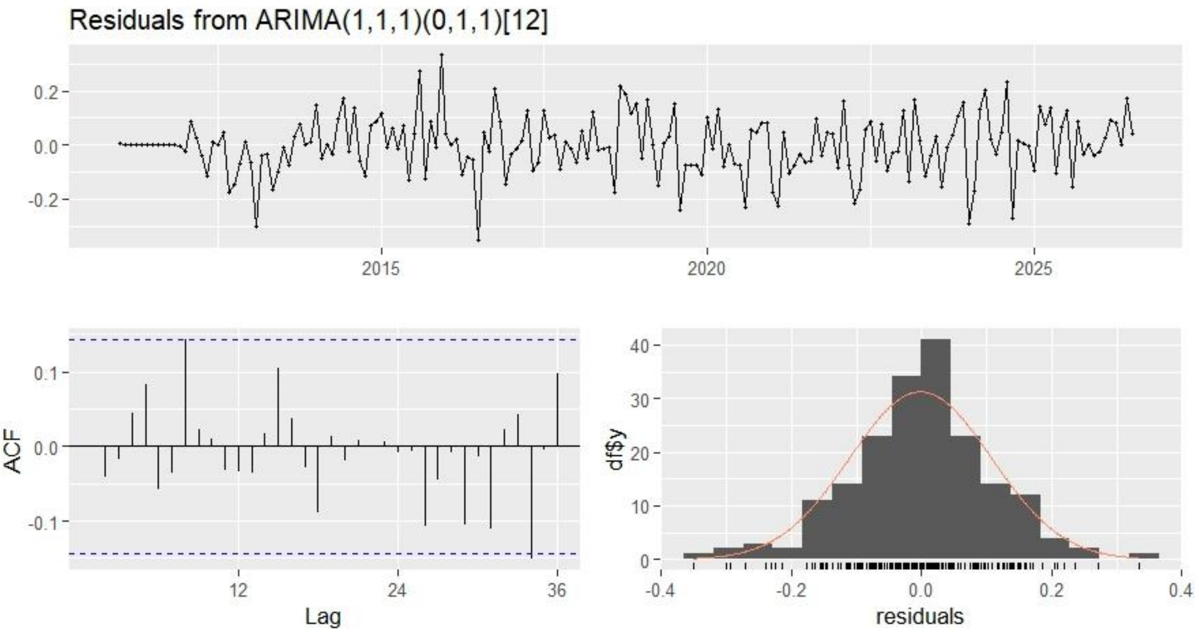


Fig.13

Diagnostics:

ARIMA(1,1,1)(0,1,1)[12]

- Coefficients:
  ar1: 0.0295 with a standard error of 0.0918.
  ma1: -0.8326 with a standard error of 0.0527.
  sma1: -0.7443 with a standard error of 0.0648.

  Particularly for ma1 and sma1, these coefficients are noteworthy because they deviate significantly from zero when compared to their corresponding standard errors.

- sigma^2: 0.01291 indicates the variance of the model residuals. The lower variance shows a better fit.
- Log Likelihood (127.55) is used along with AIC and BIC to compare models: the higher the log-likelihood, the better the model.
- AIC (-247.11): AIC is a commonly used metric to assess the quality of a statistical model. Generally, the model with the lower AIC is selected.
- BIC (-234.47): Model selection is also addressed by the BIC, which is similar to the AIC.

Ljung-Box Test:

Q* Statistic: 12.419

Degrees of Freedom: 21

p-value: 0.9276 - The absence of significant autocorrelation in the residuals at lags up to 21 is indicated by a high p-value (>0.05). This indicates that the model does not leave residual patterns, which is a feature of a well-fitted model.

From the Fig.13,

- Residuals Plot: Demonstrates residuals that are randomly arranged around zero with no evident pattern.
- ACF Plot of Residuals: This shows that there is no significant autocorrelation because almost all autocorrelations are within the confidence bounds.
- Histogram of Residuals with Density Overlay: The overlay of the normal curve indicates that the distribution closely resembles a normal distribution, indicating a good fit.

Forecast for 1 year ahead:

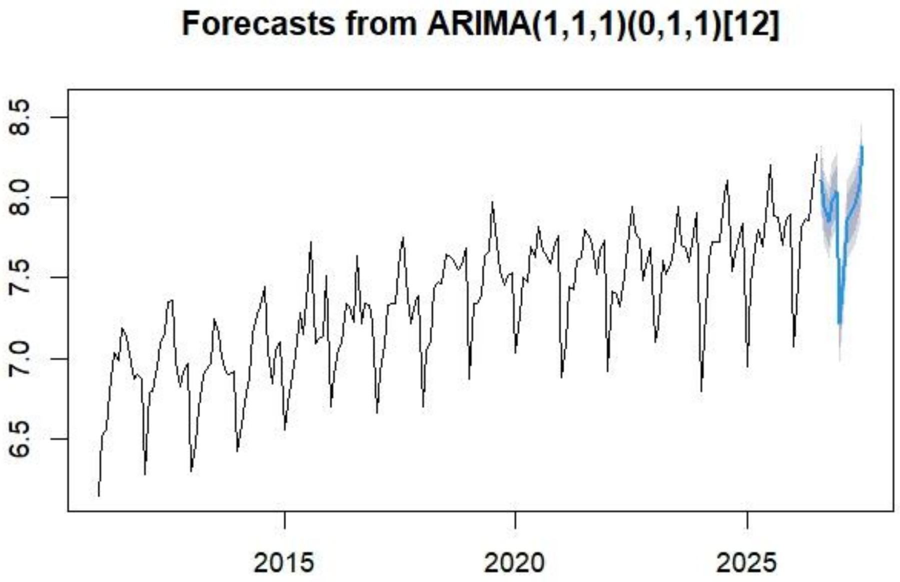


Fig.14

- The model captures the main trends without appreciable deviations, and it appears to fit the historical data well.
- The forecast covers the period into the future, starting from the last observed point.
- The forecast is shown as a line with a shaded region surrounding it, signifying the forecast uncertainty. This spread indicates the range where the future values are expected to lie.

Contrast with STL + Random walk forecasting from section 1:

STL + Random walk plot:

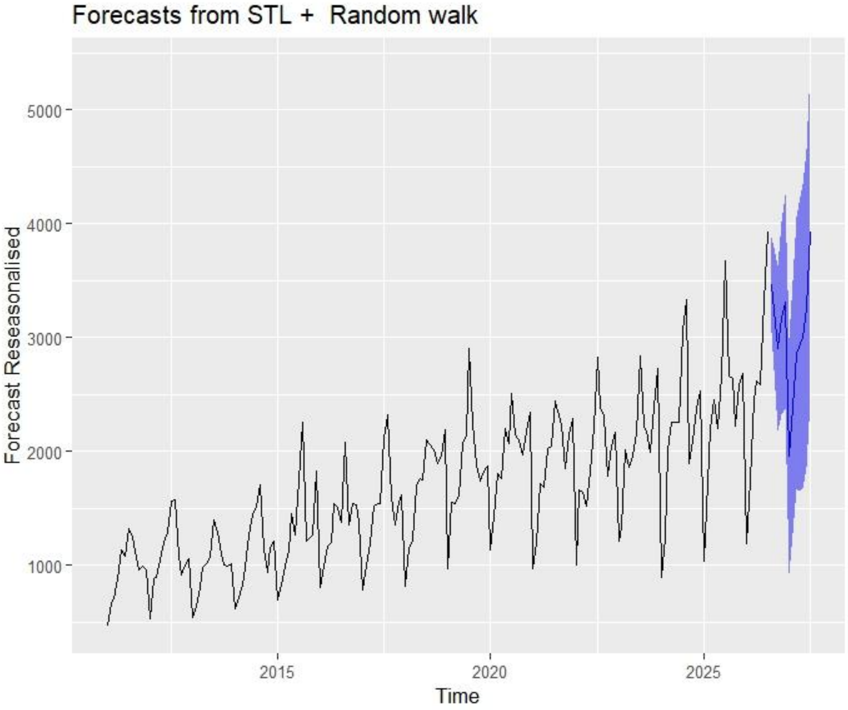


Fig.15

The blue-shaded area in the SARIMA forecast indicates a denser and narrower forecast, indicating a better prediction, whereas, the STL+Random walk forecast shows a wide forecast, indicating high uncertainty in the model's prediction.

Less volatility is seen in the SARIMA forecast, suggesting that the model can effectively capture and extend the underlying trend and seasonality. Hence, datasets where past data patterns are expected to recur or change predictably are better suited for its application.

The forecast series for the STL+Random walk demonstrates complex seasonal patterns that vary over time, making it potentially valuable in scenarios where future trends are not predicted to be very similar to past patterns.