A SARIMA-based Investigation into Room Night Occupancy Trends in Victoria

Yanru Fang

2023/12/8

Executive Summary

This research investigates the aggregate number of room nights occupied in hotels, motels, and guesthouses across Victoria, Australia, from January 1980 to June 1995. The analysis reveals a clear upward linear trend and seasonal patterns in the data, suggesting the presence of both non-stationarity and seasonality. To effectively capture these characteristics, the SARIMA(1, 1, 1)*(1, 1, 1)$_{12}$ model is identified as the most suitable for forecasting room night occupancy. Subsequent model diagnostic tests demonstrate that the residuals exhibit a symmetrical distribution yet deviate from the normality assumption. Despite this, the model upholds its forecasting capabilities, as predicted values closely align with actual data, suggesting its reliability and potential for informing local government decisions in the tourism sector.

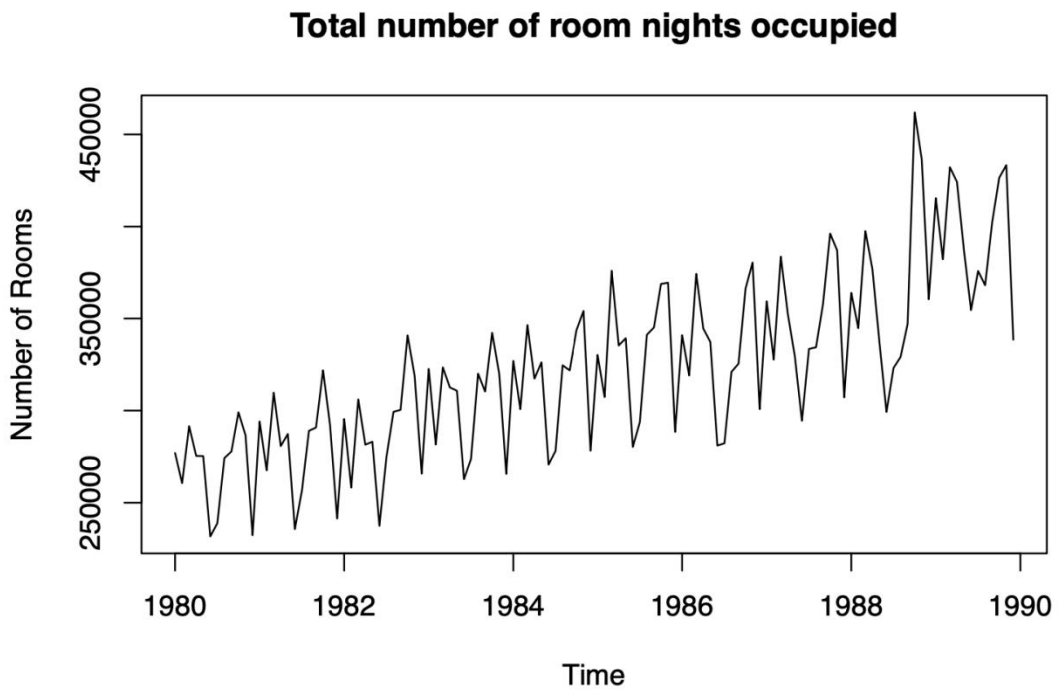
1. Introduction

Tourism plays a critical role in driving the Australian economy, exerting a substantial impact on GDP, employment, and export earnings. As one of the nation's largest industries, it employs over 600,000 individuals and stands as a cornerstone of Australia's economic advancement, playing a vital role in fostering prosperity. This article delves into an examination of the aggregate number of room nights occupied in hotels, motels, and guesthouses across Victoria, Australia, spanning from January 1980 to June 1995. With abundant tourism resources, Victoria's total hotel bookings serve as a representative indicator of the broader tourism sector. Analyzing development trends helps to comprehend Australia's tourism industry model, extract valuable insights, and offer recommendations to local governments.

The dataset, sourced from the Australian Bureau of Statistics, comprises time series data. I use the R software to perform time series modeling, uncover statistical characteristics, identify trends and seasonality within the series. Post-trend and seasonality removal, ACF and PACF guide the model selection process, with multiple candidate models compared using the AICc criterion. The ultimate choice is the SARIMA$(1, 1, 1)*(1, 1, 1)_{12}$ model. During the model diagnosis phase, the ADF test demonstrates that residuals lack relevant information, aligning with fundamental model assumptions. However, the Q-Q plot and Shapiro-Wilk test indicate non-compliance with the normality assumption. Subsequently, the model's predictions were executed, demonstrating impressive performance on the test set. The predicted data closely mirrors actual data, and the 95% prediction interval entirely encompasses real data.

2. Series Decomposition and Transformation

The dataset comprises 186 data points spanning from January 1980 to December 1989, with January 1990 serving as the demarcation point for dividing the data into training and validation sets. The data from January to Dec 1989 are incorporated into the training set. The training set accounts for 65% of the data, while the validation set comprises the remaining 35%. We employ the training set to identify an appropriate model for the sequence.



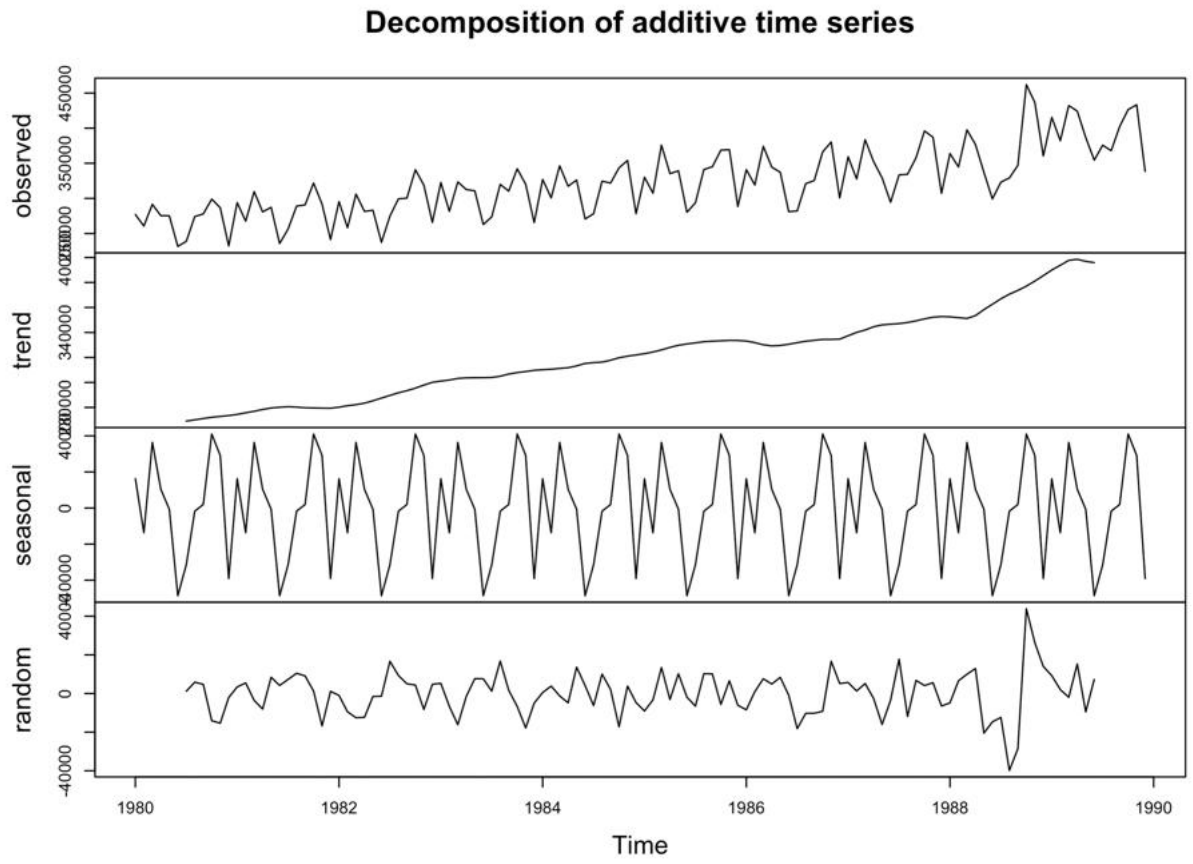


Figure 1: Time series plot and decomposition

Figure 1: depicts the time series diagram and decomposition, revealing a distinct upward linear trend and seasonal patterns. The number of room nights occupied is higher during spring and autumn months, while it is lower during summer and winter months.

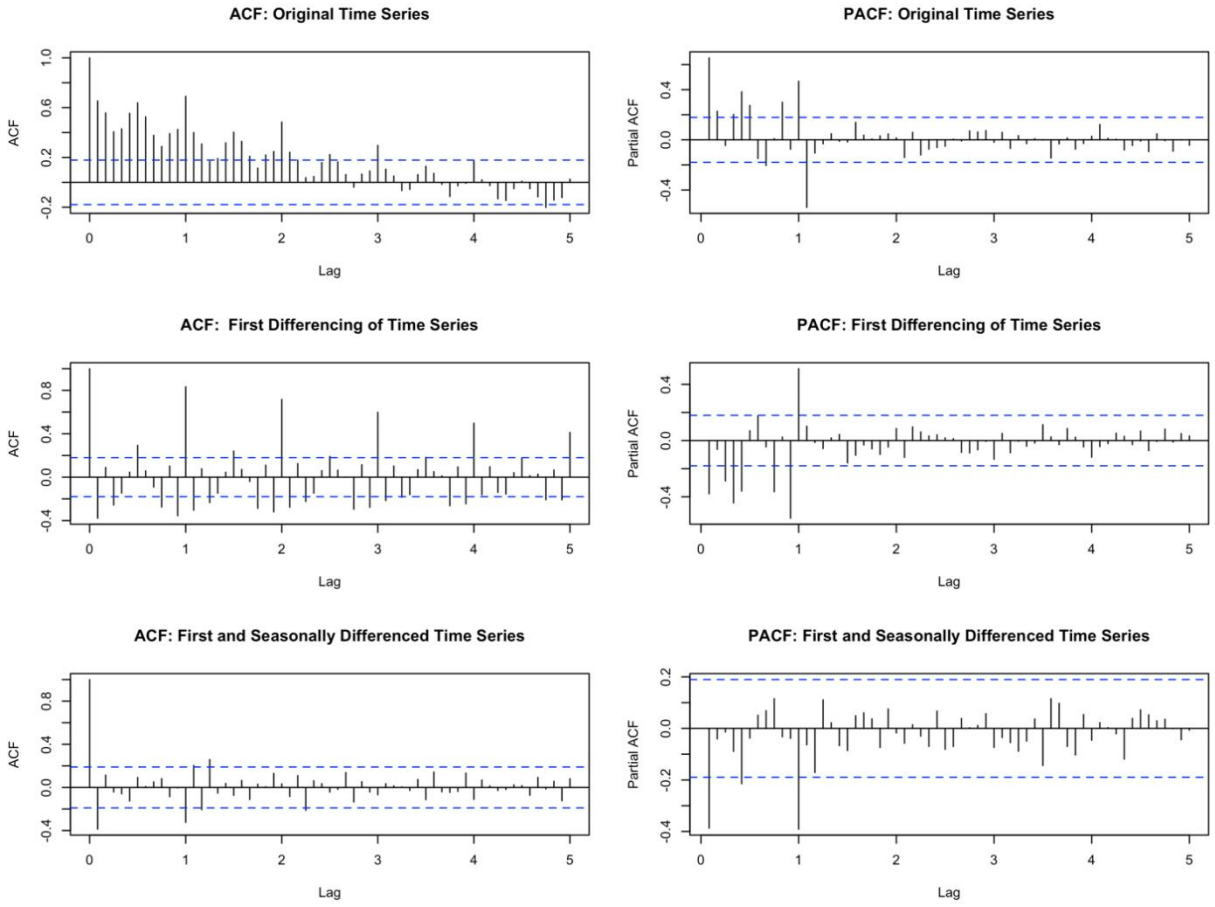


Figure 2: ACF plots and PACF plots for original series, differencing series and first and seasonality differences series

Observing the ACF and PACF plots in Figure 2 (ACF plots and PACF plots for original series, differencing series and first and seasonality differences series), we notice that the ACF of the original sequence exhibits a long tail and retains significant values at high lags, indicating non-stationarity. ACF and PACF spikes at lags 12, 24, and 36 further confirm the presence of a strong monthly seasonal effect in the series. To address the non-stationarity, we first employ first-order differencing to eliminate the trend term. This effectively reduces the tail of the autocorrelation function, but ACF and PACF still exhibit spikes at lags 12, 24, and 36. To address the seasonal effect, we subsequently apply a 12-step differencing.

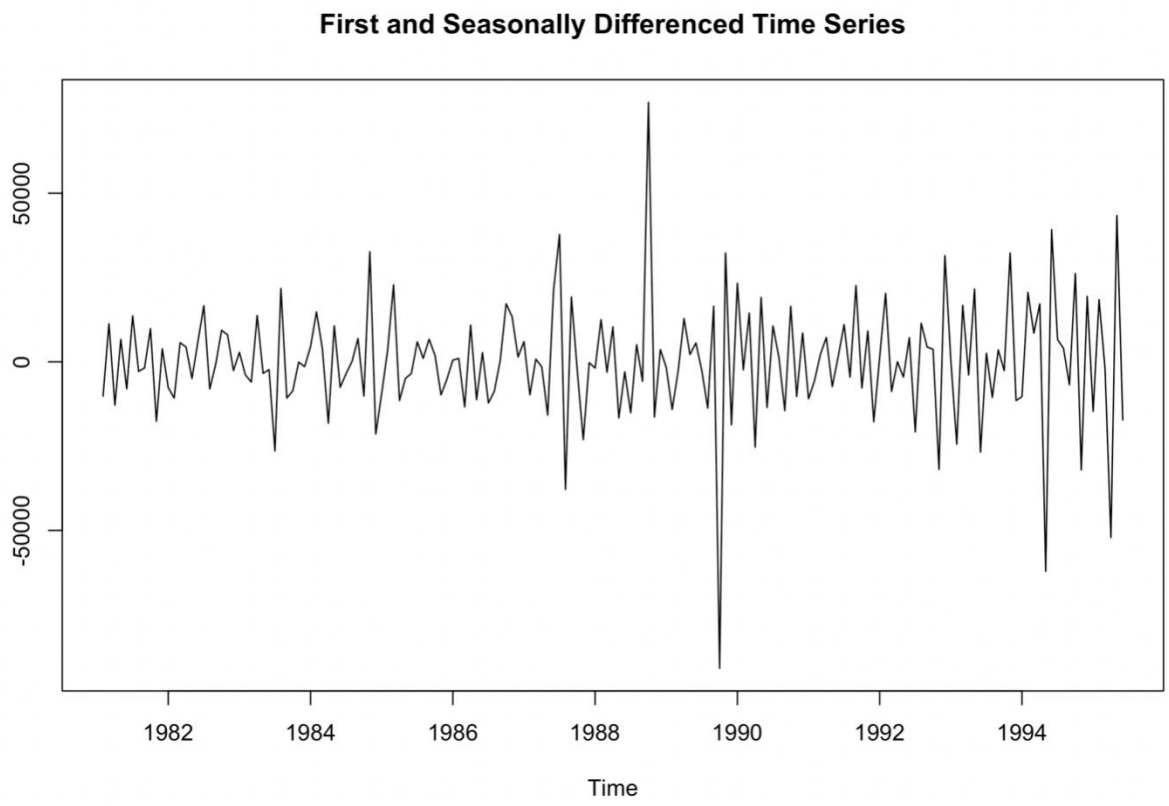


Figure 3: First and Seasonally Differenced Time Series

The resulting time series plot, shown in Figure 3 (First and Seasonally Differenced Time Series), no longer exhibits any apparent trend or seasonal effects.

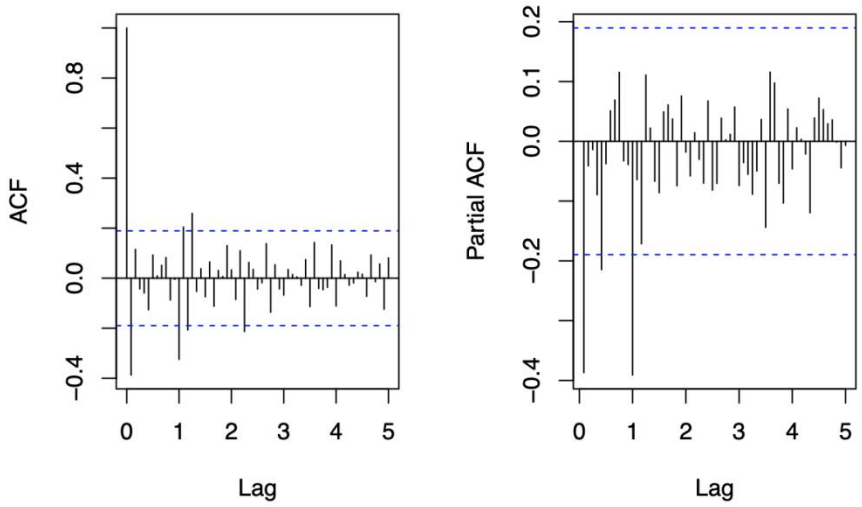
```
        Augmented Dickey-Fuller Test

data:  y_12
Dickey-Fuller = -6.4195, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Figure 4: ADF test result of First and Seasonally Differenced Time Series

The enhanced unit root test was conducted to assess the stationary of the first and seasonality differenced time series. The test statistic yielded a p-value of 0.01, leading to the rejection of the null hypothesis and the confirmation of stationary in the first and seasonality differenced time series.

3. Model Identification

After identifying the presence of trend and seasonality, I opted for a SARIMA model and proceeded to determine its optimal order based on the ACF and PACF plots of first and seasonally differenced series.



(1) Modeling the seasonal part (P, D, Q): For this part, focus on the seasonal lags h = 1s, 2s, etc.

• We applied one seasonal differencing so D = 1 at lag s = 12.

• The ACF shows a strong peak at h = 1s without peak at 2s, 3s. Then a good choice for the MA part could be Q=1

• The PACF shows two strong peaks at h = 1s without peak at 2s, 3s. Then a good choice for the AR part could be P = 1.

(2) Modeling the non-seasonal part (p , d, q): In this case focus on the within season lags, h = 1,. . . ,11.

• We applied one differencing to remove the trend: d = 1

• The ACF seems to cut off at lag 1 or 2. Then a good choice for the MA part could be q = 1 or q = 2 respectively.

• The PACF cuts off at lag h=1 or 2.

A good choice for the AR part could be p = 1 or p = 2.

Therefore, we have four candidate models now and they are

(1) SAIMAR(1, 1, 1)*$(1, 1, 1)_{12}$

(2) SAIMAR(1, 1, 2)*$(1, 1, 1)_{12}$

(3) SAIMAR(2, 1, 1)*$(1, 1, 1)_{12}$

(4) SAIMAR(2, 1, 2)*$(1, 1, 1)_{12}$

4. Model Estimation

SARIMA results

	Dependent variable:			
	Train set of Rooms			
	(1)	(2)	(3)	(4)
ar1	0.000	0.000	-0.523***	-0.523***
			(0.094)	(0.094)
ar2			-0.283***	-0.283***
			(0.103)	(0.103)
ma1	-0.553***	0.000	0.000	0.000
	(0.097)			
ma2		0.000		0.000
sar1	-0.549***	-0.445***	-0.581***	-0.581***
	(0.095)	(0.100)	(0.096)	(0.096)
sma1	0.000	0.000	0.000	0.000
Observations	107	107	107	107
Log Likelihood	-1,171.667	-1,184.551	-1,171.332	-1,171.332
$sigma^2$	181,870,569.000	235,811,247.000	179,699,223.000	179,699,134.000
AIC	21.9564	22.1785	21.9688	21.9688
AICc	21.9575	22.1789	21.9710	21.9710
Note:				*p**p***p<0.01

Table 1

Using the maximum likelihood method, we estimated the four candidate models and fixed non-significant coefficients to 0, followed by re-estimation. Table 1 presents the results. We found that SARMIA(1, 1, 1)*(1, 1, 1)$_{12}$ has the smallest AICc, indicating its superiority among the four models. And the fitted model equation is

$$(1+0.549B^{12})(1-B^{12})X_t = (1-0.5527B)Z_t$$

5. Model diagnoses and Forecast

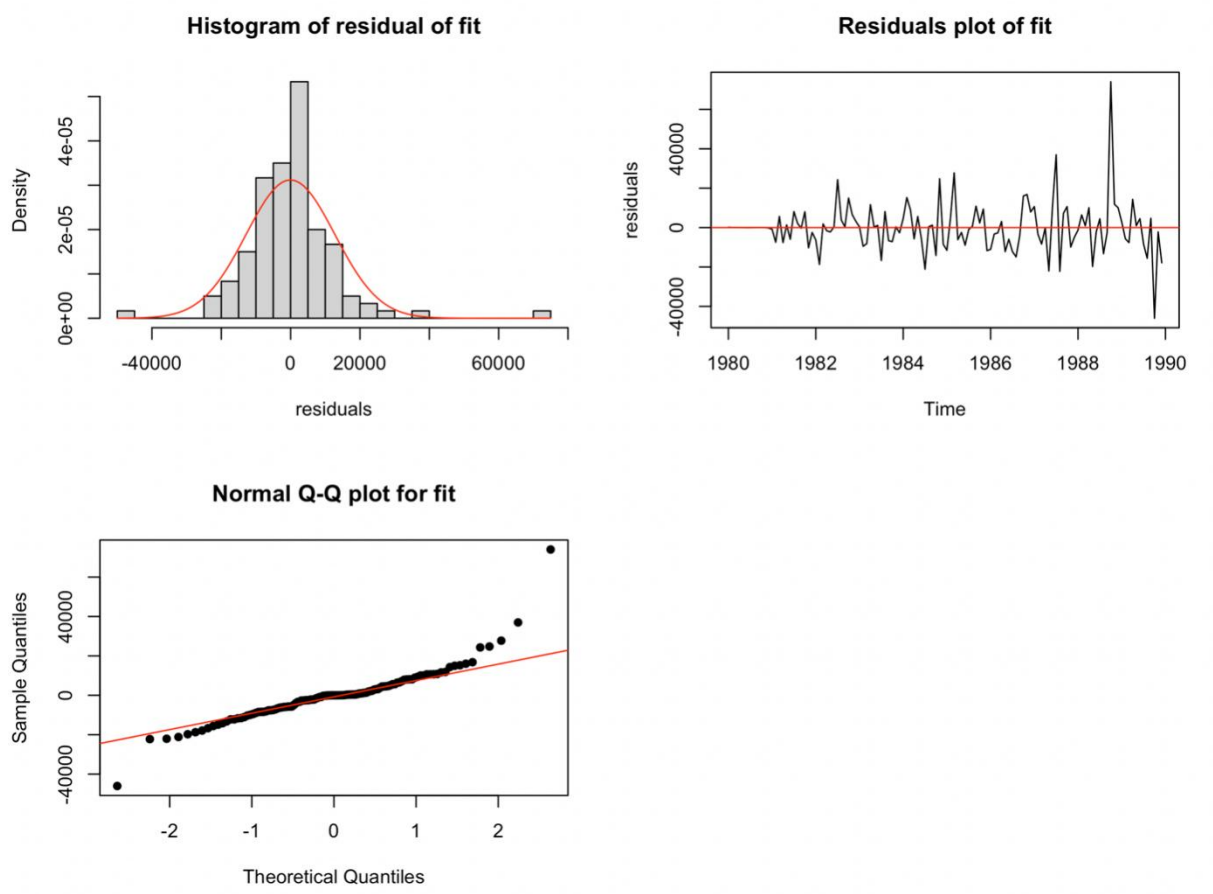


Figure 4: Model diagnoses plots

To ensure the model's effectiveness, we assessed the residuals to verify their adherence to the white noise process. Examining the residual histogram in graph 1 in Figure 4 (Model diagnoses plots) revealed a symmetrical distribution, although the peak appeared higher than that of a standard normal distribution. The Q-Q plot further indicated deviations from the straight line, particularly in the lower left and upper right corners, raising concerns about the model's normality assumption.

Then, I also check for several independence assumption.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.86892, p-value = 6.779e-09
```

The Shapiro-Wilk test confirmed our suspicions, rejecting the null hypothesis and revealing that the residuals did not conform to the normal distribution.

```
##
##  Box-Pierce test
##
## data:  residuals
## X-squared = 12.214, df = 10, p-value = 0.271
##
##  Box-Ljung test
##
## data:  residuals
## X-squared = 13.093, df = 10, p-value = 0.2185
```

To determine the adequacy of the model, we conducted the Box-Pierce and Ljung-Box tests. The p-values of 0.271 and 0.21, respectively, indicated that the residuals exhibited no significant autocorrelation.

```
##
## Call:
## ar(x = residuals, aic = TRUE, order.max = NULL, method = c("yule-walker"))
##
##
## Order selected 0  sigma^2 estimated as  163533759
```

Further, employing the yule-walker method of the ar function resulted in an automatically determined order of 0 for the residual sequence, further supporting the model's ability to capture the underlying structure without introducing spurious correlations.

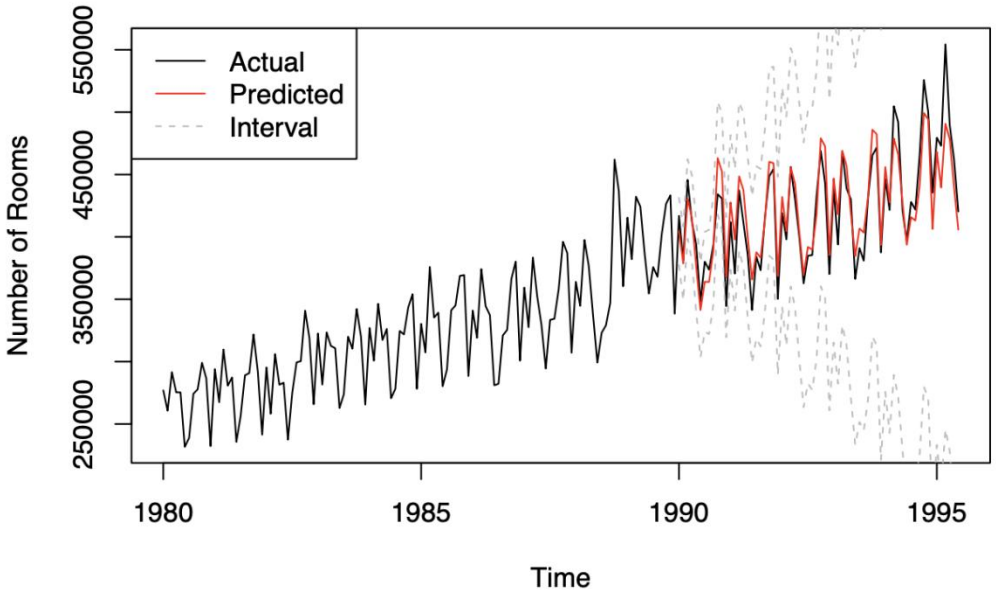


Figure 5: Prediction in test set

Finally, we validated the model's performance by applying it to the validation set and generating the corresponding 95% prediction interval. The predicted values closely mirrored the actual data, while the 95% prediction interval completely enveloped the real data, demonstrating the model's reliable forecasting capabilities.

6. Conclusion

This research provides a comprehensive analysis of room night occupancy trends in the hospitality sector across Victoria, Australia, spanning the period from January 1980 to June 1995. Employing SARIMA modeling techniques, the study successfully identifies and addresses both upward linear trends and seasonal patterns in the data. The SARIMA$(1,1,1)*(1,1,1)_{12}$ model emerges as the most suitable for forecasting room night occupancy, showcasing its effectiveness in capturing the non-stationarity and seasonality inherent in the dataset. Despite encountering deviations from the normality assumption in the residuals, the SARIMA model maintains its forecasting prowess. Model diagnostic tests, including the Box-Pierce and Ljung-Box tests, confirm the absence of significant autocorrelation in the residuals. The yule-walker method further supports the model's ability to capture underlying structures without introducing spurious correlations. The practical application of the SARIMA$(1,1,1)*(1,1,1)_{12}$ model to the validation set demonstrates its reliability and accuracy. Predicted values closely align with actual data, and the 95% prediction interval fully encompasses real data points, affirming the model's robust forecasting capabilities. These findings indicate that the SARIMA model can be a useful instrument for informing tourism-related decisions made by local governments. This includes aiding in resource allocation, policy formulation, and strategic planning.

7. Reference

[1] Hamilton, J. D. (2020). Time series analysis. Princeton university press.

[2] Harris, R., & Jago, L. (2001). Professional accreditation in the Australian tourism industry; an uncertain future. Tourism Management, 22(4), 383-390.

[3] Rob Hyndman and Yangzhouran Yang (2018). tsdl: Time Series Data Library.

8. Appendix

```
library(tsdl)
# Select tourism data with frequency of 12
tourism <- subset(tsdl, 12, "Transport and tourism")
# Select total number of room nights occupied
rooms_ts <- ts(tourism[[2]][, 1], start = 1980, frequency = 12)
# Split into training dataset and test dataset
rooms_train <- ts(rooms_ts[1:120], start = 1980, frequency = 12)
rooms_test <- ts(rooms_ts[121:186], start = 1992, frequency = 12)

# Time series plot
ts.plot(rooms_train, ylab = "Number of Rooms")
title("Total number of room nights occupied")

# Component decomposition
plot(decompose(rooms_train))

par(mfrow = c(3, 2))
# Get differential series
y_1 <- diff(rooms_train, 1)
y_12 <- diff(y_1, 12)

# Plot ACFs and PACFs
acf(rooms_train, lag.max = 60, main = "ACF: Original Time Series")
```

```
pacf(rooms_train, lag.max = 60, main = "PACF: Original Time Series")
acf(y_1, lag.max = 60, main = "ACF: First Differencing of Time Series")
pacf(y_1, lag.max = 60, main = "PACF: First Differencing of Time Series")
acf(y_12, lag.max = 60, main = "ACF: First and Seasonally Differenced Time Series")
pacf(y_12, lag.max = 60, main = "PACF: First and Seasonally Differenced Time Series")

par(mfcol=c(1,1)) plot(y_12, ylab = "", main = "First and Seasonally Differenced Time Series")

# Stationary test
library(tseries)
adf.test(y_12)

par(mfrow = c(1, 2))
acf(y_12, lag.max = 60, main = "")
pacf(y_12, lag.max = 60, main = "")

# Model estimations
library(astsa)
fit1 <- sarima(xdata = rooms_train, p = 1, d =1, q =1, P = 1, D = 1, Q = 1, S = 12, details = F)
fit1.1 <- sarima(xdata = rooms_train, p = 1, d =1, q =1, P = 1, D = 1, Q = 1, S = 12, details = F,
fixed=c(0,NA,NA,0))
fit2 <- sarima(xdata = rooms_train, p = 1, d =1, q =2, P = 1, D = 1, Q = 1, S = 12, details = F)
fit2.1 <- sarima(xdata = rooms_train, p = 1, d =1, q =2, P = 1, D = 1, Q = 1, S = 12, details = F,
fixed=c(0,0,0,NA,0))
fit3 <- sarima(xdata = rooms_train, p = 2, d =1, q =1, P = 1, D = 1, Q = 1, S = 12, details = F)
fit3.1 <- sarima(xdata = rooms_train, p = 2, d =1, q =1, P = 1, D = 1, Q = 1, S = 12, details = F,
fixed=c(NA,NA,0,NA,0))
fit4 <- sarima(xdata = rooms_train, p = 2, d =1, q =2, P = 1, D = 1, Q = 1, S = 12, details = F)
fit4.1 <- sarima(xdata = rooms_train, p = 2, d =1, q =2, P = 1, D = 1, Q = 1, S = 12, details = F,
fixed=c(NA,NA,0,0,NA,0))
```

```
# Fit the final model
fit <- arima(rooms_train, order=c(1,1,1), seasonal = list(order = c(1,1,1) ,period = 12) ,
fixed=c(0,NA,NA,0), method="ML")

par(mfrow = c(2,2))
# Obtain the residual
residuals <- residuals(fit)

# Histogram of residual
hist(residuals, breaks = 30, freq = F, main = "Histogram of residual of fit")
curve(dnorm(x, mean(residuals), sd(residuals)), add = T, col = "red")

# Time series of residuals
plot.ts(residuals, main = "Residuals plot of fit")
abline(h = 0, col = "red")

# Q-Q plot of residuals
qqnorm(residuals, pch = 16, main = "Normal Q-Q plot for fit")
qqline(residuals, col = "red")

# Test for normality
shapiro.test(residuals)

# White noise test
Box.test(residuals, lag = 10, type = c("Box-Pierce"))
Box.test(residuals, lag = 10, type = c("Ljung-Box"))
ar(residuals, aic = TRUE, order.max = NULL, method = c("yule-walker"))

# Prediction
library(forecast)
n.head = length(rooms_test)
```

```
pred.tr <- predict(fit, n.ahead = n.head)
U.tr= pred.tr$pred + 2*pred.tr$se
L.tr= pred.tr$pred - 2*pred.tr$se
par(mfrow = c(1, 1))
ts.plot(rooms_ts, ylab = "Number of Rooms")
lines(pred.tr$pred, col = "red")
lines(U.tr, col="grey", lty="dashed")
lines(L.tr, col="grey", lty="dashed")
legend("topleft", legend=c("Actual", "Predicted","Interval"),
       col=c("black", "red", "grey"), lty=c(1, 1, 2))
```