Tracy Zhu, Yucheng Zhao

Professor Aue

STA 137

07 June 2023

Analyzing Trends in Travel Interest to Cancun:
A Time-Series Approach Utilizing Google Trends Data

Abstract

This research aims to forecast the volume of the US tourists traveling to Cancun, Mexico, utilizing an Autoregressive Moving Average (ARMA) model. By analyzing Google Trends data from 2006 to 2019, this study seeks to develop a model that can predict fluctuations in tourism. Initial observations indicated a clear upward trend in the data, with obvious seasonality. After log transformation to ensure stable variance, the data was examined for trend and seasonal components. The resultant ARMA (1,1) model demonstrated a significant fit, suggesting that it can effectively predict tourist volume. This study contributes valuable insights into tourist behavior and the predictive power of search engine data, with implications for tourism industry stakeholders in Cancun.

Introduction

Cancun, Mexico, has been a prominent tourist destination for many years. The influx of tourists has significant economic impacts on the region, and understanding trends in this visitation is of great importance to a variety of stakeholders.

This research aims to construct such a model by employing an ARMA (1,1) approach. We leveraged Google Trends data to gauge the popularity of Cancun as a travel destination. Google Trends provides an unparalleled opportunity to examine the internet search behavior of millions of users, which has increasingly been used as a proxy for various societal and economic trends. In this case, we use it to track the volume of tourists intending to visit Cancun.

By tracking the keyword "flight to Cancun" from January 2006 to December 2019, we have identified patterns in the data. Given the global effects of the COVID-19 pandemic on tourism, the data after 2019 were treated as outliers and removed from the sample. The aim is to capture a "normal" operating environment for Cancun tourism. The data has been log-transformed to manage variance and carefully scrutinized to identify any trend or seasonal components.

This study will present the ARMA (1,1) model developed from this data and discuss its potential applications and limitations. Through this research, we aim to offer a predictive tool that can be used to understand better the trends in the tourism industry of Cancun, Mexico, and potentially assist stakeholders in planning and decision-making processes.

Data description

We utilized the search data of the keyword “flight to Cancun” in Google Trends to refer to the number of US tourists traveling to Cancun, as most US tourists always need to search for and buy flight tickets to Cancun before traveling. As a result, this keyword precisely reflected the number of tourists. In the following analysis, the total number of tourists traveling to Cancun from January 2006 to May 2023 is used. Considering the impact of COVID-19 on tourism, the data after 2019 can be removed as outliers, and finally, we selected from January 2006 to December 2019 as sample data. The sample data is based on a monthly basis, and the number of data observations included is 168. The variable of interest was the total number of tourists going to Cancun.

Data analysis

Data cleaning

We found no missing values or outliers in the original dataset.

Initial data transformations

According to the time series plot of the original data, we can find that the total number of tourists to Cancun has an obvious upward trend and a seasonal trend. The log-transformed time series plot shows that the time series has a clear upward trend with a seasonal trend (fig. 1). But the log-transformed time series fluctuate significantly less than the original time series, so it is reasonable to believe that the log-transformed data have more stable variance. Therefore, we select the log-transformed data as the sample data.

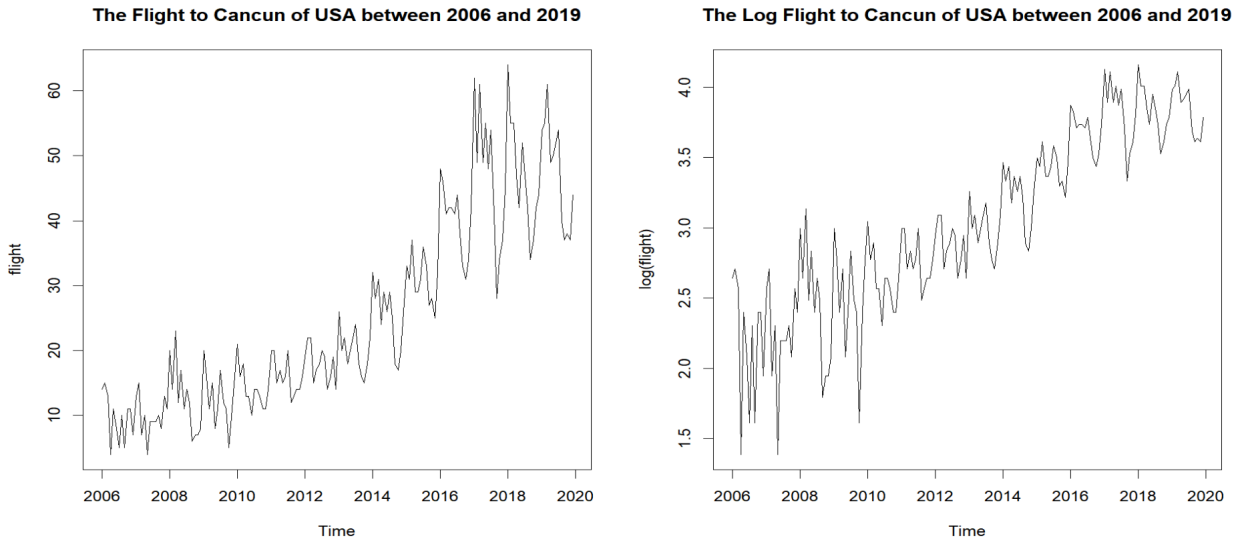

Fig. 1. Original Data Transformed

Analyzing the "smooth" component

In order to find out the trend component, we fit two models with log (flight) as the dependent variable with t and t^2 as the independent variable. The fitting results are as follows:

```
Call:
lm(formula = log(flight) ~ t)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96633 -0.18017 -0.00037  0.20322  0.78354

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.0338979  0.0445880   45.62   <2e-16 ***
t           0.0117797  0.0004577   25.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2877 on 166 degrees of freedom
Multiple R-squared:  0.7996,    Adjusted R-squared:  0.7984
F-statistic: 662.5 on 1 and 166 DF,  p-value: < 2.2e-16
```

According to the fitting results, it can be found that the independent variable t is statistically significant, while t^2 is not statistically significant. So we finally choose the

fitting line with only the independent variable t, and add the fitting line to the time series map, which can find that the trend component can be well reflected (fig. 2).

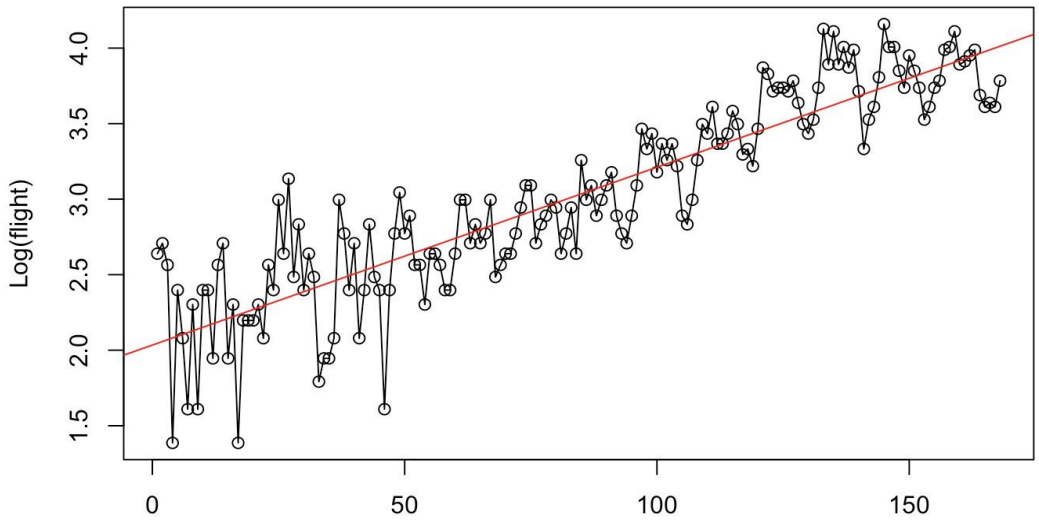

Fig. 2. Fitted line

To find the seasonal component, we eliminate the trend component and use the small trend method to estimate the seasonal component (fig. 3) (Tsafack).

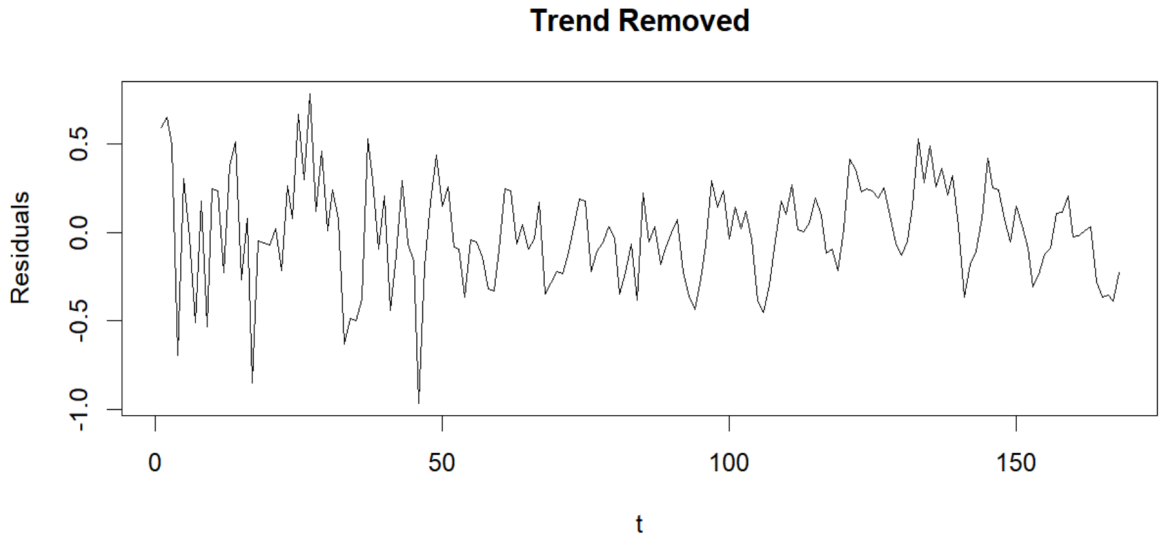

Fig. 3. Trend removed

Then, we get the final seasonal component (fig. 4).

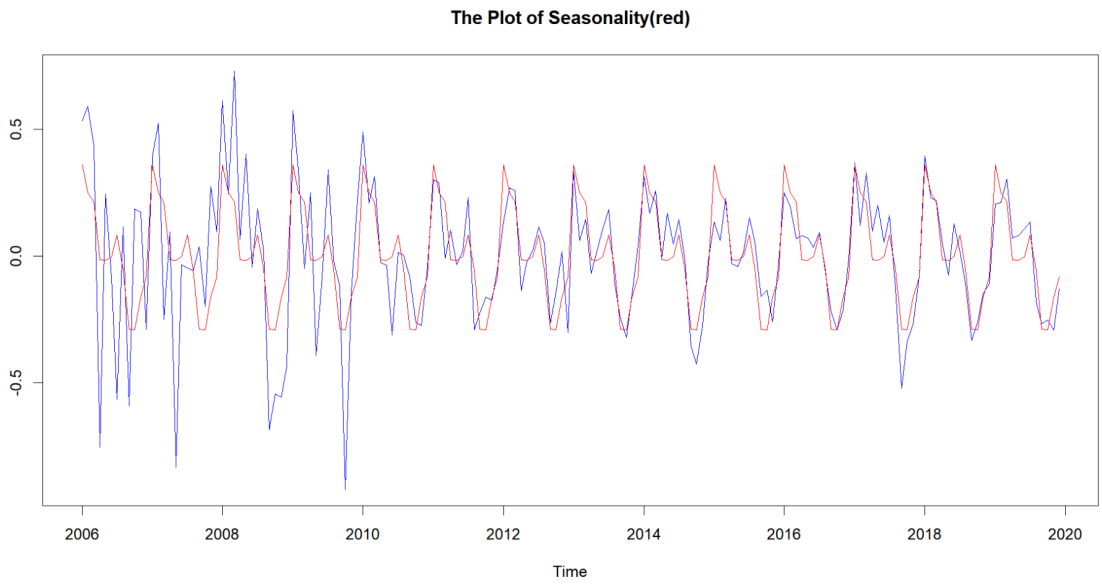

Fig. 4. Seasonal component

Analyzing the residuals

The time series plot shows that the original sequence has obvious seasonal factors. After extracting the trend and seasonal components, we need to eliminate trend and seasonal components, and the final residual plot is as follows (fig. 5):

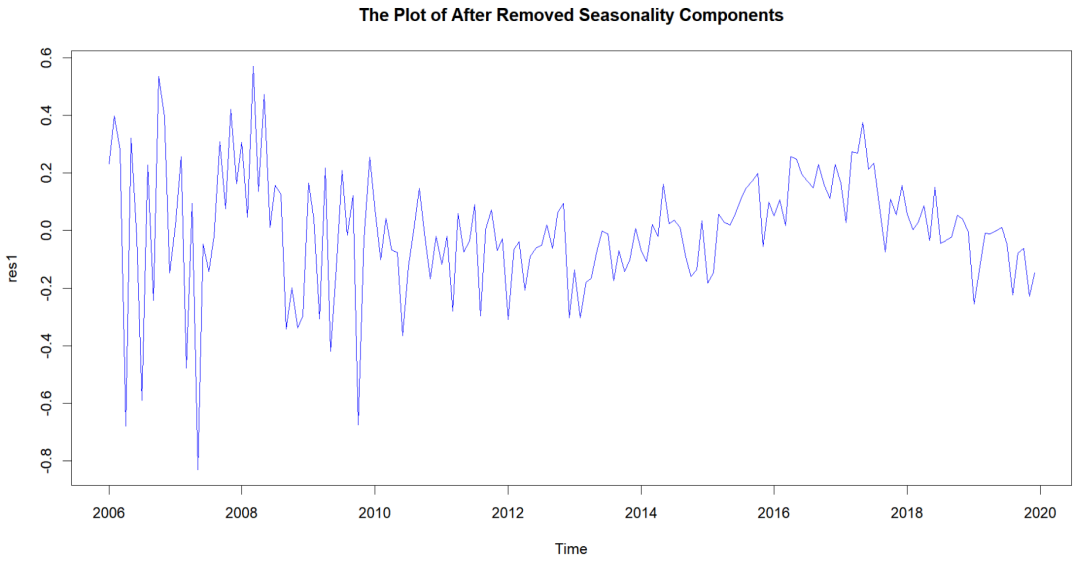

Fig. 5. Final residual

According to the final residual time series plot, it seems to see that the residual satisfies the white noise and there is no remain trend, but we need to ensure that the residual is stable, using Box-Ljung test (Quantstart), the test results are:

```
        Box-Ljung test

data:  res1
X-squared = 2.8686, df = 1, p-value = 0.09032
```

The test results show that the p-value is 0.09032, which is greater than 0.05 when given the significance level of 5%, so we cannot reject that the residuals satisfy the white noise sequence. Further testing of residual normality was performed by Q-Q plot (fig. 6) and showed that the residual basically satisfied the normality assumption.

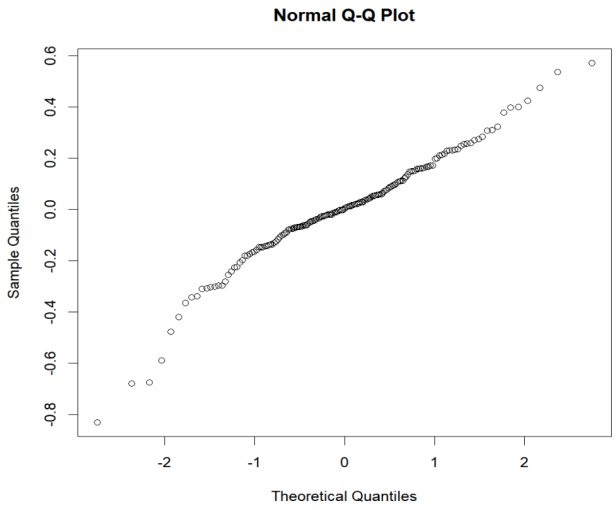

Fig. 6. Normal Q-Q plot

Analyzing the "rough" component

ARMA Model Fitting

The ACF and PACF plot (fig. 7) from the stationary residue time series obtained above.

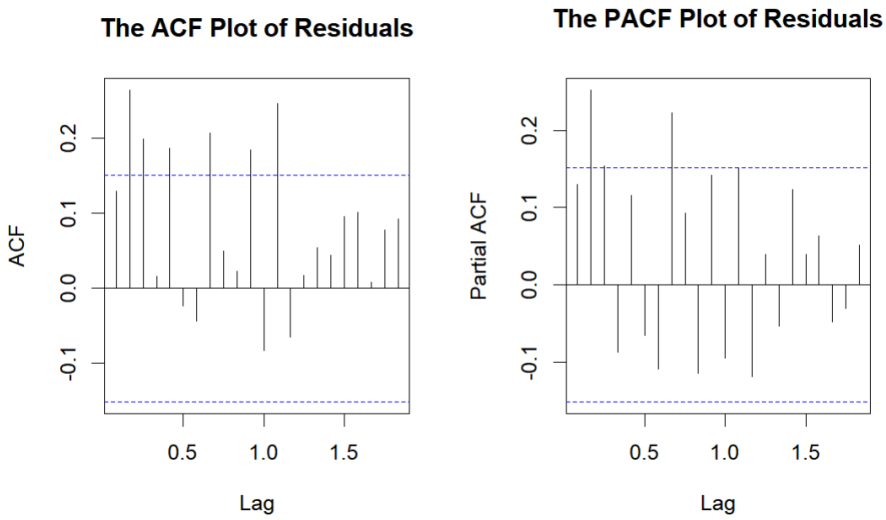

Fig. 7. ACF and PACF

The ACF plot shows that the sequence has 1 order censoring, while the PACF plot shows that the sequence has 1 order trailing, so we decided to use ARMA (1,1) model for fitting (Parra).

```
Call:
arima(x = z, order = c(1, 0, 1), include.mean = T, method = "CSS-ML")

Coefficients:
         ar1      ma1  intercept
      0.8762  -0.7399     0.0000
s.e.  0.1015   0.1440     0.0321

sigma^2 estimated as 0.04116:  log likelihood = 29.53,  aic = -53.06
```

Next, the residuals of the resulting ARMA (1,1) model need to be tested. We use Box-Ljung test again:

```
        Box-Ljung test

data:  wn
X-squared = 3.609, df = 2, p-value = 0.1646
```

According to the test results, it can be found that the p-value is 0.1646, which is greater than 0.05 when given a significance level of 5%, so we cannot reject that the residuals are stable. We map the ARMA(1,1) model to the previously obtained rough residuals (fig. 8). We can see from the comparison diagram that the ARMA (1,1) model fit highly coincides with the rough residuals obtained previously, so the matching rate is very high.

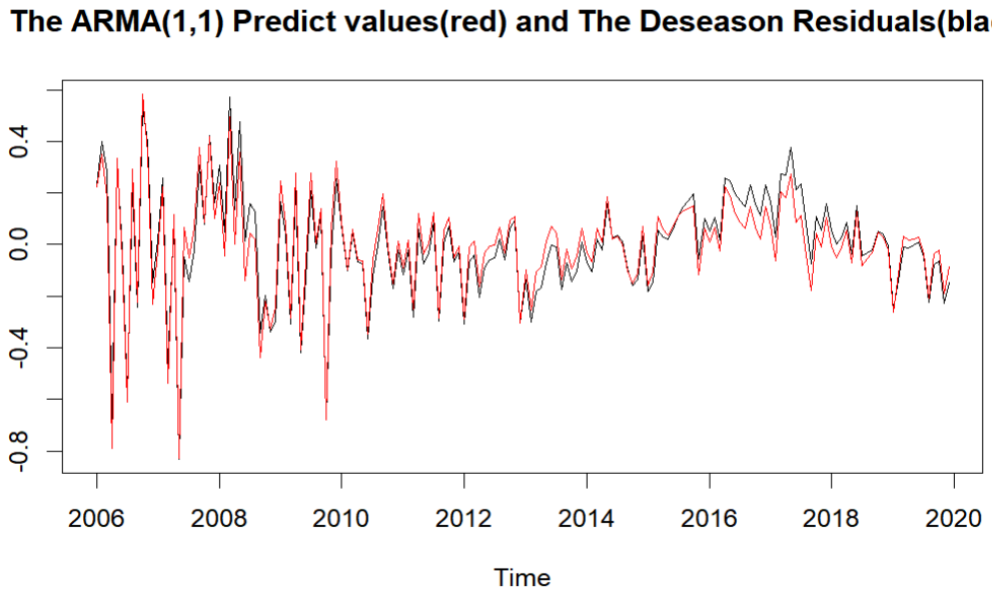

Fig. 8. ARMA(1,1) and Residuals

Predict future values

Using the resulting model, excluding the impact of COVID-19, we forecast the next eleven years' total number of trips to Cancun, which is from January 2020 to December 2030. First, we obtain the ARMA model residuals, then we obtain the trend and seasonal components and finally add the three parts to obtain the time series prediction value.

Spectral Analysis

There is a peak at the frequency of 1/12, which means that there is a yearly seasonality as we use monthly data.

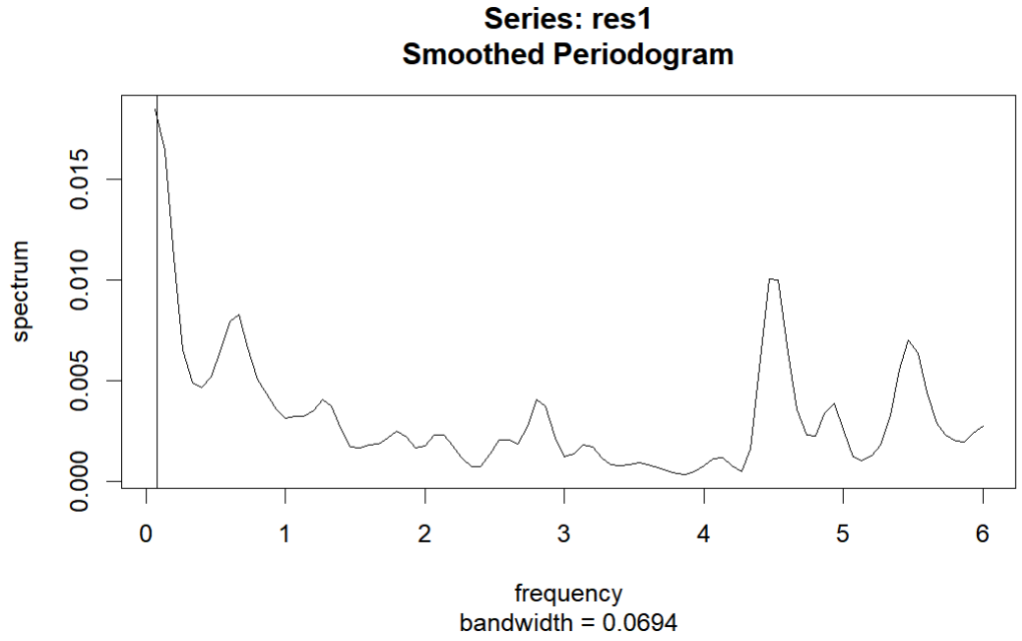

Fig. 9. Smoothed Periodogram

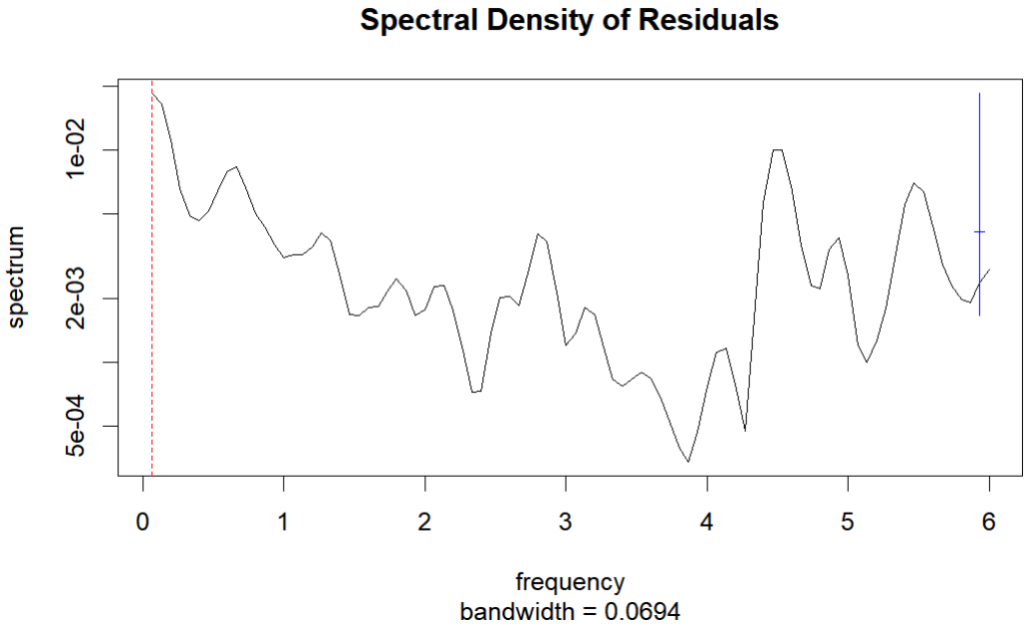

Fig. 10. Spectral Density

Discussion

In conclusion, the log-treated time series fluctuates significantly less strong than the original time series, so it is reasonable to believe that the log-transformed data have a more stable variance. The trend-fitting results show a significant linear trend in the time series. The residual has a white noise sequence, which is smooth. The predicted value obtained by the ARMA (1,1) model fit highly matches the residues obtained by the smooth part. The final forecast result of the model shows that when we eliminate the impact of COVID-19 factors, the total number of tourists to Cancun, Mexico, will gradually increase in 2020. According to the data analysis in the article, the log transformation obviously outperforms the original data on the original data transformation.

However, the sample data we selected may have some deviations because the global outbreak of COVID-19 in December 2019 will significantly affect the development of global tourism, so there are some defects in the sample data. This model is only able to predict the normal situation.

What's more, although the square term of t is not significant in the fitting to the trend component, the higher order term may be significant, and for convenience, we performed only a linear trend fit. Maybe we should try to fit more models. We should also try to fit more types of ARMA models after getting more fit types.

Then, we should do a more detailed spectral analysis if we have more time.

Conclusions

The analysis results show that the total number of tourists to Cancun, Mexico, showed a linear upward trend from 2006 to 2019, and the total number of tourists showed a seasonal trend. ARMA (1,1) is a very good prediction of the number of tourists traveling to Cancun, Mexico. The forecast shows that the number of tourists to Cancun, Mexico, will gradually increase if we ignore COVID-19.

Reference

"Flight to Cancun." trends.google.com/trends/explore?date=all&geo=US&q=flight%20to%20cancun&hl=en. Accessed 1 Jun. 23.

"Time Series Analysis (Aue)." *LibreTexts*, 23 Feb. 2021, stats.libretexts.org/Bookshelves/Advanced_Statistics/Time_Series_Analysis_(Aue). Accessed 12 Jun. 2023.

"Autoregressive Moving Average ARMA(P, Q) Models for Time Series Analysis - Part 3." *Quantstart*, www.quantstart.com/articles/Autoregressive-Moving-Average-ARMA-p-q-Models-for-Time-Series-Analysis-Part-3/#:~:text=Choosing%20the%20Best%20ARMA(p,achieved%2C%20for%20particular%20values%20of%20. Accessed 1 Jun. 2023.

"A Complete Introduction To Time Series Analysis (with R):: Model Selection for ARMA(P,Q)." *Hair Parra*, 29 Jan. 2021, medium.com/analytics-vidhya/a-complete-introduction-to-time-series-analysis-with-r-model-selection-for-arma-p-q-ebc338e6d159. Accessed 1 Jun. 23.

"ARMA Models with R: The Ultimate Practical Guide with Bitcoin Data." *Idriss Tsafack*, 8 Dec. 2020, www.idrisstsafack.com/post/arma-models-with-r-the-ultimate-practical-guide-with-bitcoin-data. Accessed 1 Jun. 23.

All lecture notes, discussion materials, and homework solutions.

Appendix

```
library(zoo)
library(forecast)
library(quadprog)
library(quantmod)
require(forecast)
library(stats)
flight <- read.csv("flight_to_cancun.csv")
#only keeps the data from 2016 to 2019
flight <- flight[,1]
flight <- as.numeric(flight[c(26:193)])

#transform into a time series object
flight=ts(flight,start=c(2006, 1),frequency = 12)
```

```
n = length(flight)
t <- 1:n
#save 2 plots in the same file
par(mfrow=c(1,2))
ts.plot(flight, main="The Flight to Cancun of USA between 2006 and 2019")
ts.plot(log(flight), main="The Log Flight to Cancun of USA between 2006 and 2019")
fit <- lm(log(flight) ~ t)
summary(fit)
log_flight <- plot(x = t, y = log(flight), type="o", ylab="Log(flight)")
log_flight
#fitting line
abline(fit, col = "red")

t2 <- t * t
fit1 <- lm(log(flight) ~ t+t2 )
summary(fit1)
yhat = fitted(fit)
plot(x = t, log(flight), type="o", ylab="Log(flight)")
abline(fit, col = "red")
y = residuals(fit)
plot(t,y, type="l", main="Trend Removed", ylab="Residuals")

# Small trend method
# a vector of the average residual values for each year
m_j1=tapply(log(flight)-yhat,floor(time(log(flight))),mean)
m_j1=ts(rep(m_j1, 12),start=2006,frequency = 12)
# residuals after removing trend, mean of residuals
ts.plot(log(flight)-yhat,m_j1,col=c("black","red"))
```

```
# residuals removed both trend and mean
ts.plot(log(flight)-yhat-m_j1,col="blue")

#average seasonality
s_k1=tapply(log(flight)-yhat-m_j1,cycle(log(flight)),mean)
s_k1=ts(rep(s_k1,times=14),start=2006,frequency = 12)
#deseasonalized residuals and average seasonality
ts.plot(log(flight)-yhat-m_j1,s_k1,col=c("blue","red"),main="The Plot of Seasonality(red)")
#residuals after removed the trend and seasonality
res1=log(flight)-yhat-s_k1
ts.plot(res1,col="blue",main="The Plot of After Removed Seasonality Components")
#check if the residual satisfies white noise
Box.test(res1, lag = 1, type = "Ljung-Box")
qqnorm(res1)
z = res1
par(mfrow=c(1,2))
acf(z,main="The ACF Plot of Residuals")
pacf(z,main="The PACF Plot of Residuals")
#ARMA model
fit3 <- arima(z, order=c(1,0,1), include.mean=T, method='CSS-ML')
summary(fit3)
wn = resid(fit3)
#plots the ACF, PACF of residuals from the ARMA model
acf(wn, lag.max=10)
pacf(wn, lag.max=10)
#test the ARMA model
Box.test(wn, lag = 2, type = "Ljung-Box")
```

```
ts.plot(res1,wn,col=c("black","red"),main="The ARMA(1,1) Predict values(red) and The Deseason Residuals(black)")

#forecast the next 10 month

fc = forecast(wn, h=132, level = .95)

plot(fc)

fc

#fit a linear reg model using flight data with y

fit4 = lm(y~.,data=flight)

t_new = 169:300

#predicted trend using t_new

trend_fc = fit$coefficients[1]+fit$coefficients[2]*t_new

trend_fc

#predicted seasonality

season_fc = fit4$fitted.values[1:10]+fc$mean


#get the predicted number of flights to Cancun from 1/2020 to 10/2020

x_hat = season_fc+trend_fc

x_hat

flight = exp(x_hat)

flight


# Spectral Density

spec_obj = spec.pgram(res1, spans = c(3,3), log = "no")

abline(v=1/12)

plot(spec_obj, main="Spectral Density of Residuals")

# Peak Frequency

peak_frequency = spec_obj$freq[which.max(spec_obj$spec)]

abline(v=peak_frequency, lty=2, col="red")
```