TSDL - Electricity production Analysis

1234

2021/11/23

Abstract

In this project, I am going to use the time series analysis methods from PSTAT 174 class on the data set in the tsdl library. The goal of this project is use the SARIMA model to fit a model of election usage in Australia, and finally forecast future consumption. First, I will use plot to show the data, determine if Box-Cox transformation is needed, difference is required to make it equal variance and stationary. Then I would inspect the sample ACF, PACF plot to propose candidate models. AICc will be calculated for those candidate models and the model with smallest AICc will be selected the best model. Then I will check the residual of the model, inspecting the sample ACF, PACF, some independence tests, normality tests, check characteristic roots to prove that the residual is quite alike white noise, and the model is staionary and invertible. Last, I will proceed to forecast the next three year's data, and check the true data with the 95% confidence interval.

Introduction and data description

Analysis the electricity production data could help us predict industry activity and economy behaviors. Without electricity, production has to be suspended, and electricity data should be seasonal to the econnomy circle. In this project I am going to analysis the time series data of the power production in Australia.

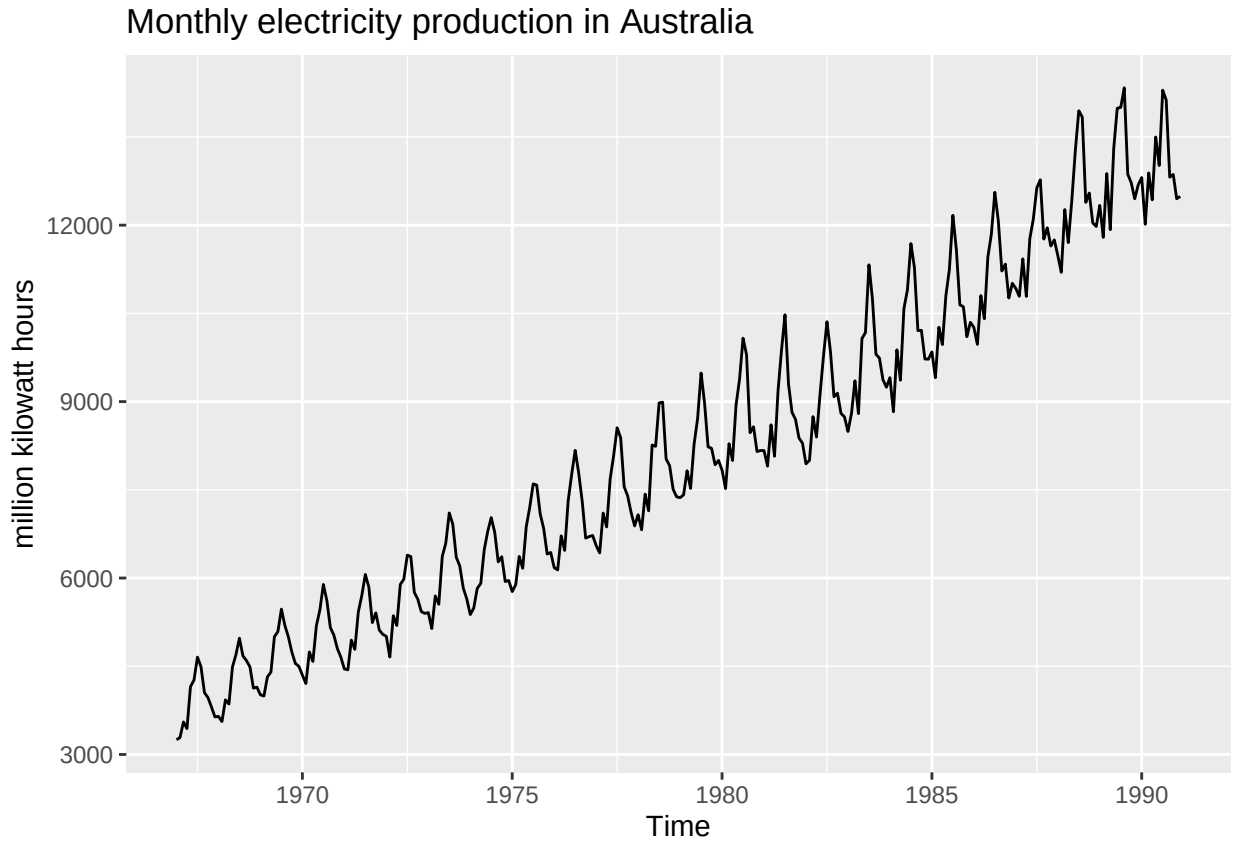
The data comes from the tsdl dataset with index 122. The source is from the Australian Bureau of Statistics, and the data is about the Monthly electricity production in Australia: million kilowatt hours. Jan 1967 - Aug 1995.

In this project I will use the data after 1991 as the test set, the data before 1991 as the train set to train the SARIMA model.

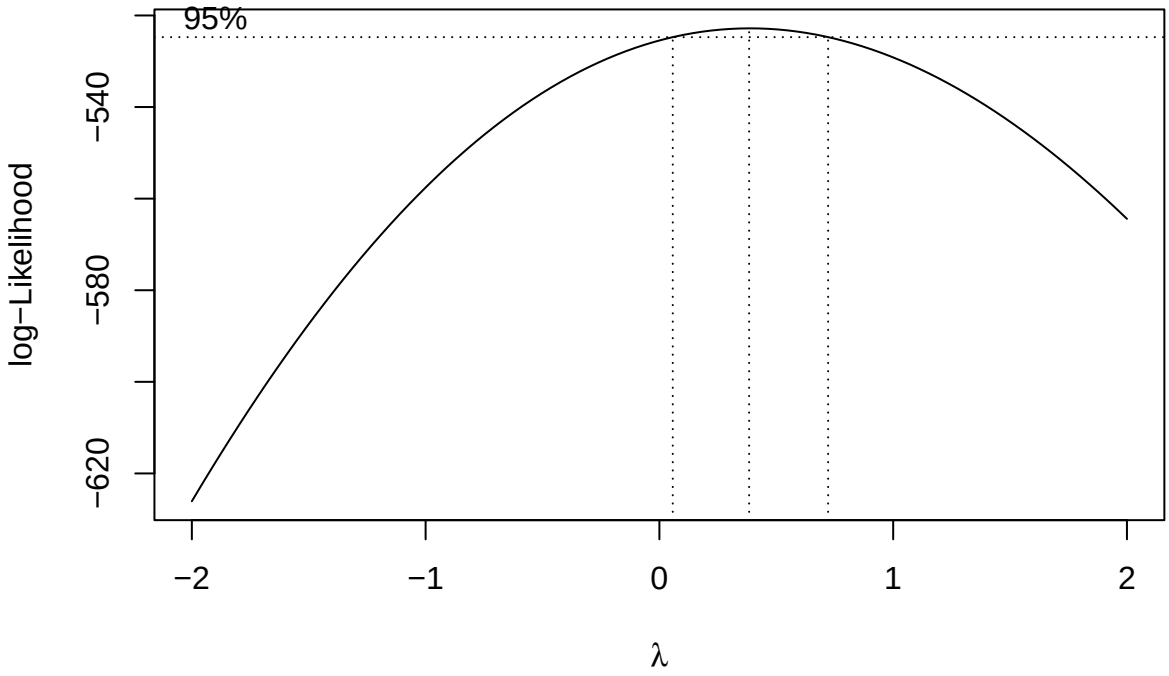
```
## $tsp
## [1] 1956.000 1995.583   12.000
##
## $class
## [1] "ts"
##
## $source
## [1] "Australian Bureau of Statistics"
##
## $description
## [1] "Monthly electricity production in Australia: million kilowatt hours. Jan 1956 - Aug 1995"
##
## $subject
## [1] "Production"
```

Exploratory analysis

First plot the train data here:

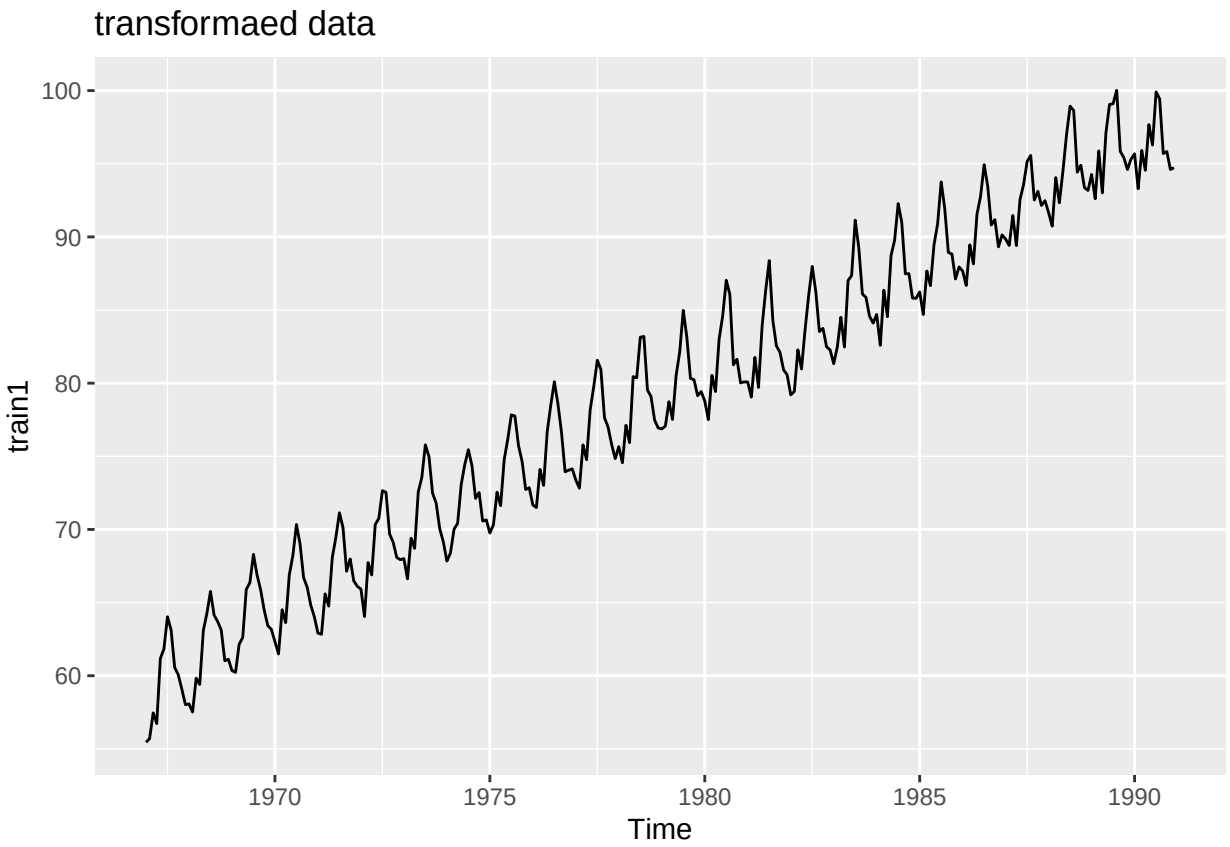


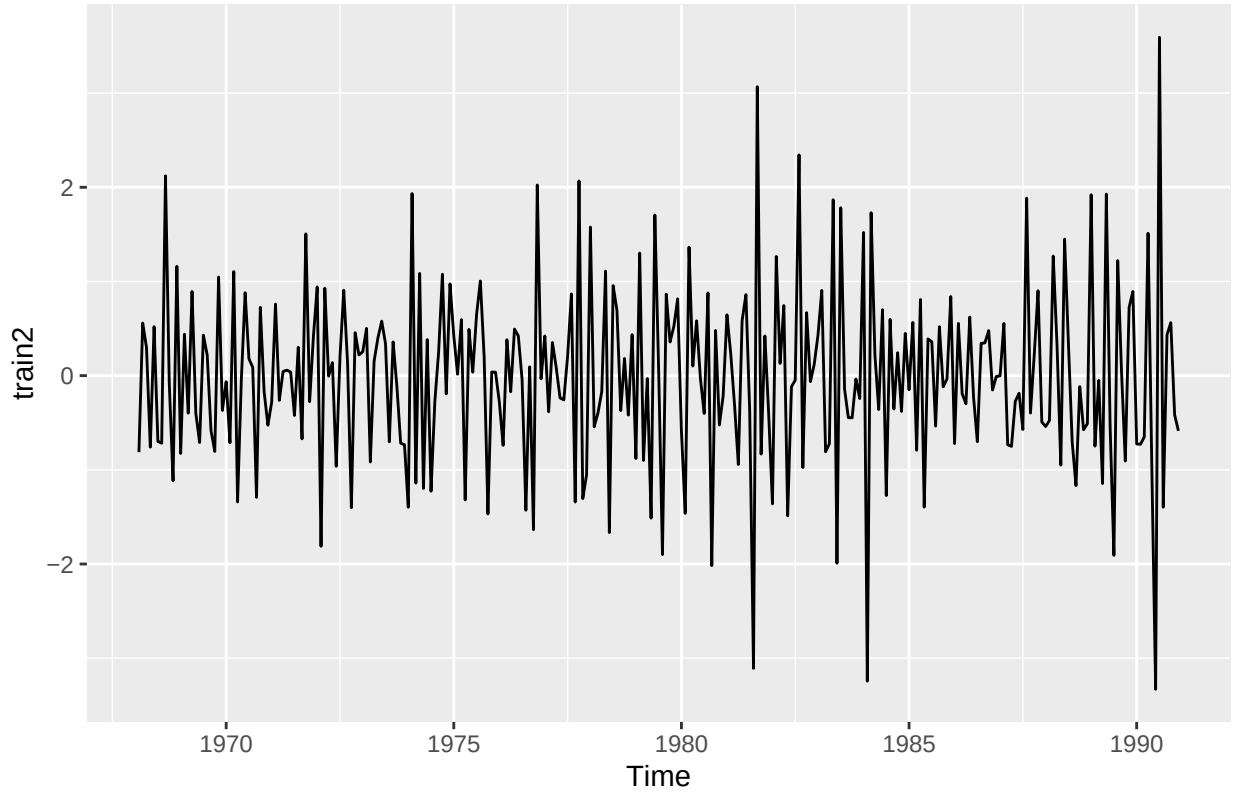
Here I observe that the time sequence has increasing trend and clear seasonal effect, also the variance seems to be increasing. The first thing need to do is to use Box Cox transformation to stablize the variance.



As the 95% CI of the optimal lambda in the Box Cox transformation does not contain 1, so we use the optimal lambda = 0.384 in R as the box cox transformation parameter.

Then to make the series stationary, we need to make 1 lag 1 difference and 1 lag 12 difference. After that, I plot the time series and also do an Augmented Dickey-Fuller Test. The null hypothesis is the series is not stationary, and the alternative hypothesis is that the series is stationary. As the p value is smaller than 0.01, I believe it is stationary after lag 1 and lag 12 difference.


transformaed data
train1
100
90
80
70
60
1970
1975
1980
1985
1990
Time

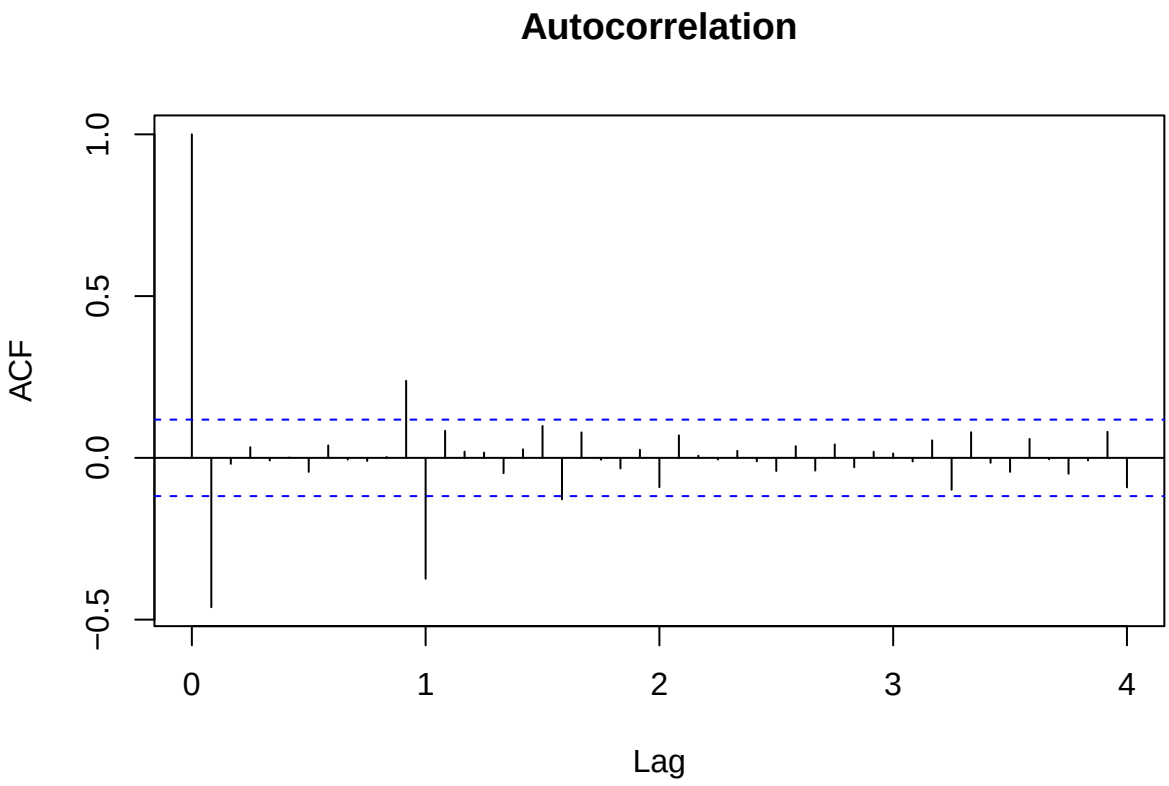


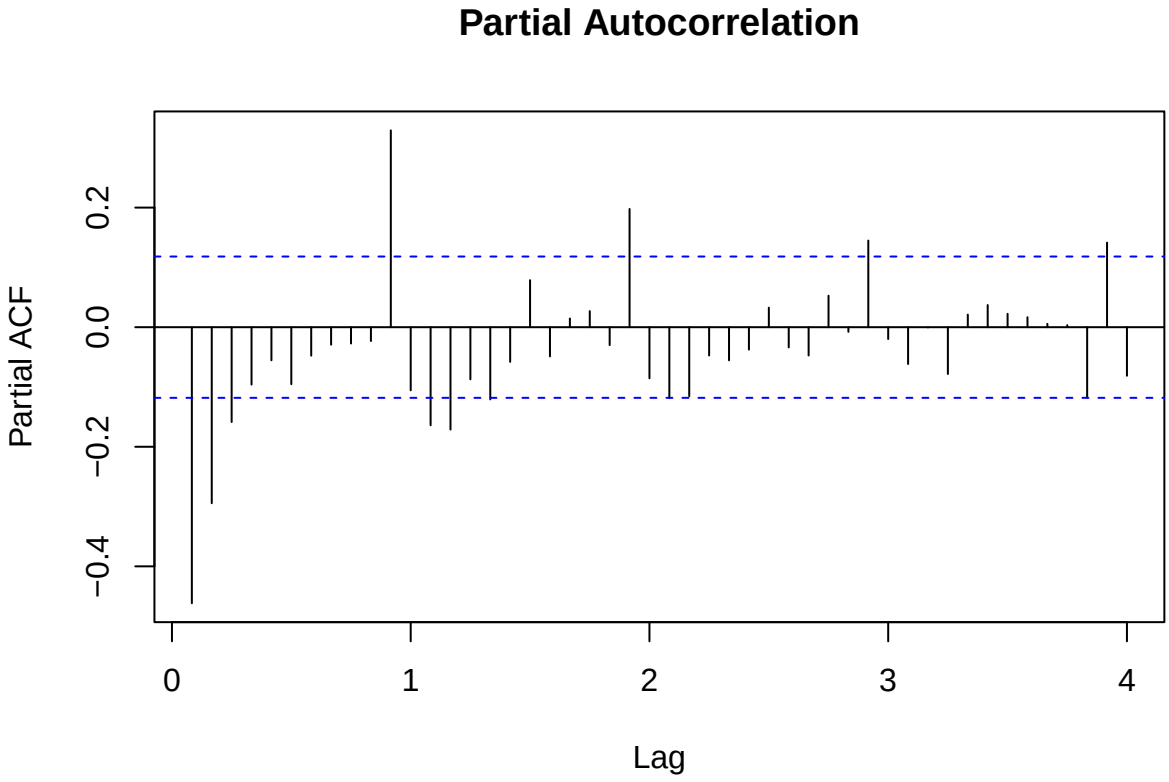

```
##
##  Augmented Dickey-Fuller Test
##
## data:  train2
## Dickey-Fuller = -8.5713, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

Now after all of these transformations, the series seems stationary and ready for fitting ARIMA models.

Model selection

I use the sample ACF, PACF to find the candidate models:


Autocorrelation
ACF
1.0
0.5
0.0
−0.5
0
1
2
3
4
Lag



From the ACF plot, in lag 1-11, the only significant lag is lag 1, so the MA part should be 1. For the seasonal part, there is significant spike at lag 12 , so candidate SMA parts is SMA(1).

From the PACF plot, the lag tails off from 1 to 11, and the seasonal lag also tails off, so I think the ARIMA part should be $ARIMA(0,1,1)$. For the seasonal part, the seasonal spikes also tails off, to the SAR part I would choose 0, thus the candidate model is $SARIMA(0,1,1)(0,1,1)_{12}$

To determine the best model, I fit all possible model with the possible parameters ranged from 0 to 2. And all those model AICc are provided in the table below: (NA means the model can't be fitted in R)

Table 1: AICcs of the candidate models

	(0,1,0)_12	(1,1,0)_12	(2,1,0)_12	(0,1,1)_12	(1,1,1)_12	(2,1,1)_12	(0,1,2)_12	(1,1,2)_12	(2,1,2)_12
ARIMA(0,1,0)	756.4451	710.4500	686.2969	654.3213	655.0345	655.8451	654.7939	655.5816	657.6565
ARIMA(1,1,0)	692.5274	642.4392	616.2481	592.8409	594.5300	595.5070	594.4564	595.6609	597.6420
ARIMA(2,1,0)	669.5391	618.7940	589.5020	571.0786	573.0278	573.3655	572.9932	573.8157	575.2835
ARIMA(3,1,0)	664.6732	614.1858	583.5428	565.6985	567.7268	567.5019	567.7063	568.3651	569.1739
ARIMA(0,1,1)	655.2385	604.7334	574.7872	555.8706	557.8272	557.9482	557.7972	558.3366	559.5832
ARIMA(1,1,1)	657.2787	606.5984	576.4052	557.9293	559.8779	NA	559.8659	560.4037	NA
ARIMA(2,1,1)	659.2953	608.5447	578.0740	559.9751	561.9662	562.0892	561.9364	562.4776	563.6960
ARIMA(3,1,1)	655.3105	610.5255	580.1384	561.9138	563.9020	564.0636	563.8667	564.4133	NA

From the table we find that our candidate model do have the smallest AICc with 555.87, so we will choose that as the best model. The model summary is here:

```
## Series: train1
```

```
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##           ma1     sma1
##       -0.6714  -0.7363
## s.e.   0.0453   0.0474
##
## sigma^2 estimated as 0.4201:  log likelihood=-274.89
## AIC=555.78   AICc=555.87   BIC=566.63
##
## Training set error measures:
##                       ME      RMSE       MAE        MPE     MAPE      MASE
## Training set -0.04200297 0.6310284 0.4569974 -0.0580518 0.577431 0.2740608
##                      ACF1
## Training set -0.00235966
```

As both ma1 and sma1 parameter are outside of the 2 standard deviation, so they are both significant.

Check the model ARIMA(0,1,1)(1,1,1):

```
## Series: train1
## ARIMA(0,1,1)(1,1,1)[12]
##
## Coefficients:
##           ma1    sar1     sma1
##       -0.6706  0.0277  -0.7499
## s.e.   0.0456  0.0864   0.0621
##
## sigma^2 estimated as 0.4214:  log likelihood=-274.84
## AIC=557.68   AICc=557.83   BIC=572.15
##
## Training set error measures:
##                       ME      RMSE       MAE         MPE      MAPE      MASE
## Training set -0.04210113 0.6308602 0.4559093 -0.05813829 0.5761214 0.2734082
##                       ACF1
## Training set -0.0008510136
```

The coefficient of sar1 is not significant here, so we should not add the SAR part here.

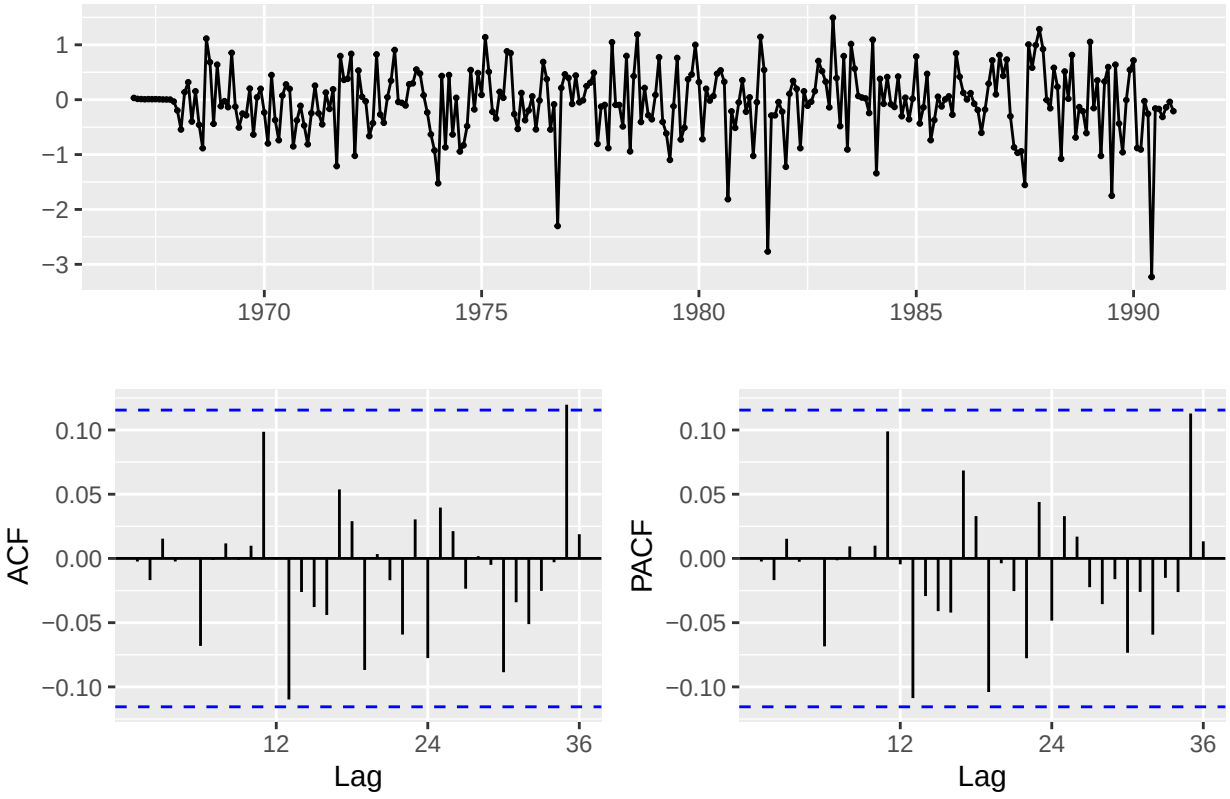
So Our model is: (After box cox transformation of lambda 0.384‘)

$$X_n - X_{n-1} - X_{n-12} + X_{n-13} = Z_n - 0.6714Z_{n-1} - 0.7363Z_{n-12} + 0.4944Z_{n-13}, Z_n \overset{iid}{\sim} N[0, 0.4201]$$

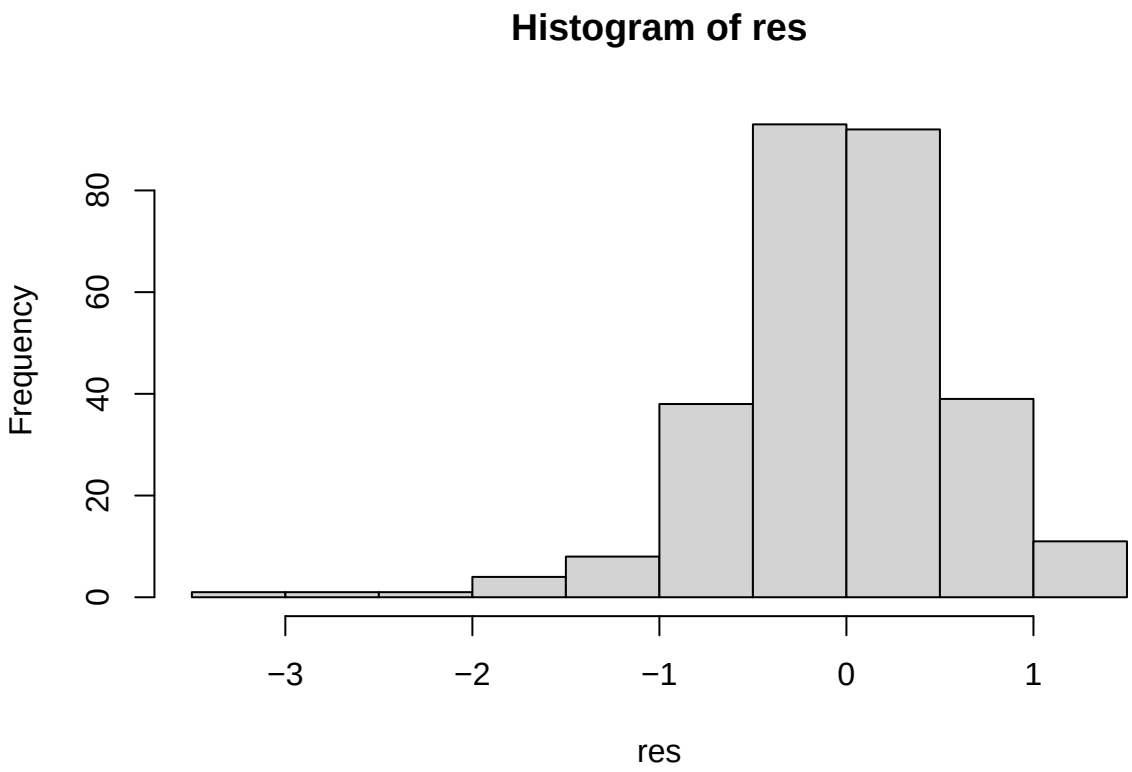
Model diagnostic

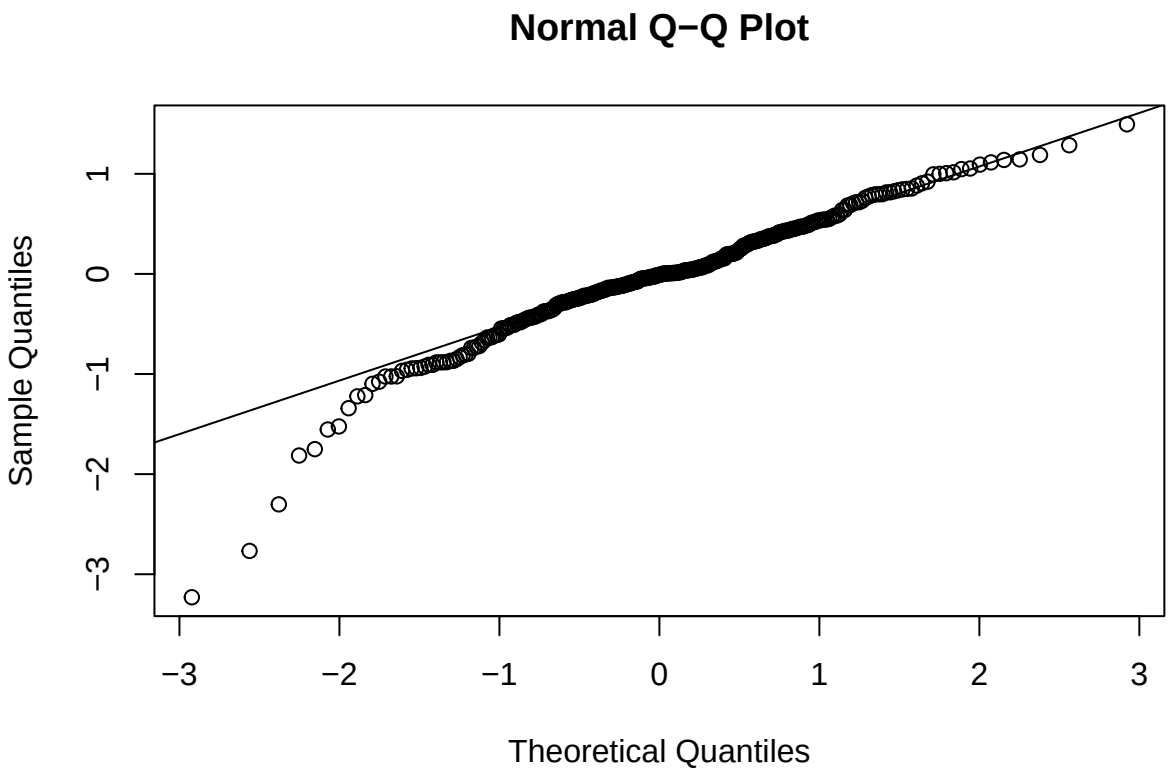
In this part I am going to check the model with residual, and test whether they are like white noises.

First check the sample ACF and the sample PACF, and there is no significant spikes. (except some marginal spikes) Possible a more complicated model other than SARIMA would better fit this series.



Then I check the residual about normality as below:


Histogram of res
Frequency
0
20
40
60
80
−3
−2
−1
0
1
res




```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.95217, p-value = 4.307e-08
```

Though the p value of the Shapiro test was smaller than 0.05, and it shows the residual is not normal, but when we check the histogram and the qqplot, the non normality is because some extreme residuals with small values (<-2). If we only consider the residuals that are between -1 0 and 1, they are approximately normal. So I think the normality is not a problem of the model even the Shapiro Wilks test is not pass, but it is because of the data with some unusual bumps. And I think as only a few fraction of points are off the line, it won't influence the confidence interval much.

- Model residual independence:

Use Box - Pierce test, Ljung - Box test and McLeod - Li test. The training data set has 288 observations, so the degree of freedom for the first two tests are 17 - 2 = 15, the degree of freedom for the McLeod - Li test is 17. I put the p value below:

tests	p.value	df
Box-Pierce	0.8304607	15
Ljung-Box	0.7989196	15
McLeod-Li	0.4865023	17

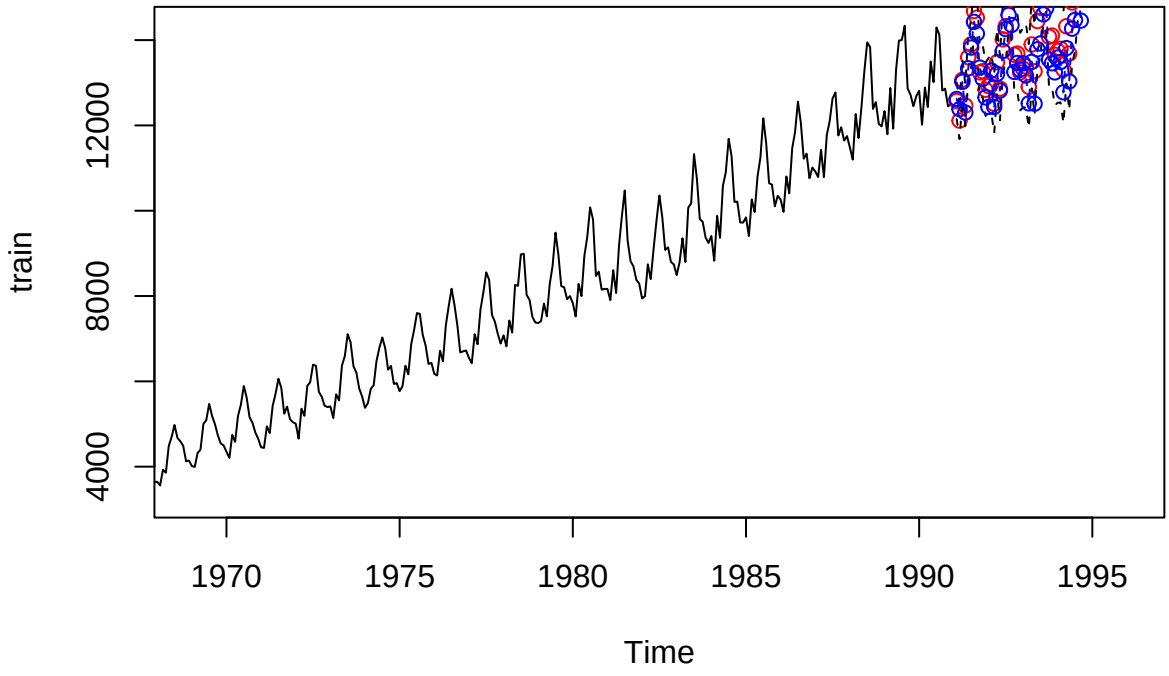
From the table, all of the p values are bigger than 0.05, so all independence tests are satisfied, I can believe the residual is independent.

For stationary and invertible, we check the characeteristic roots:

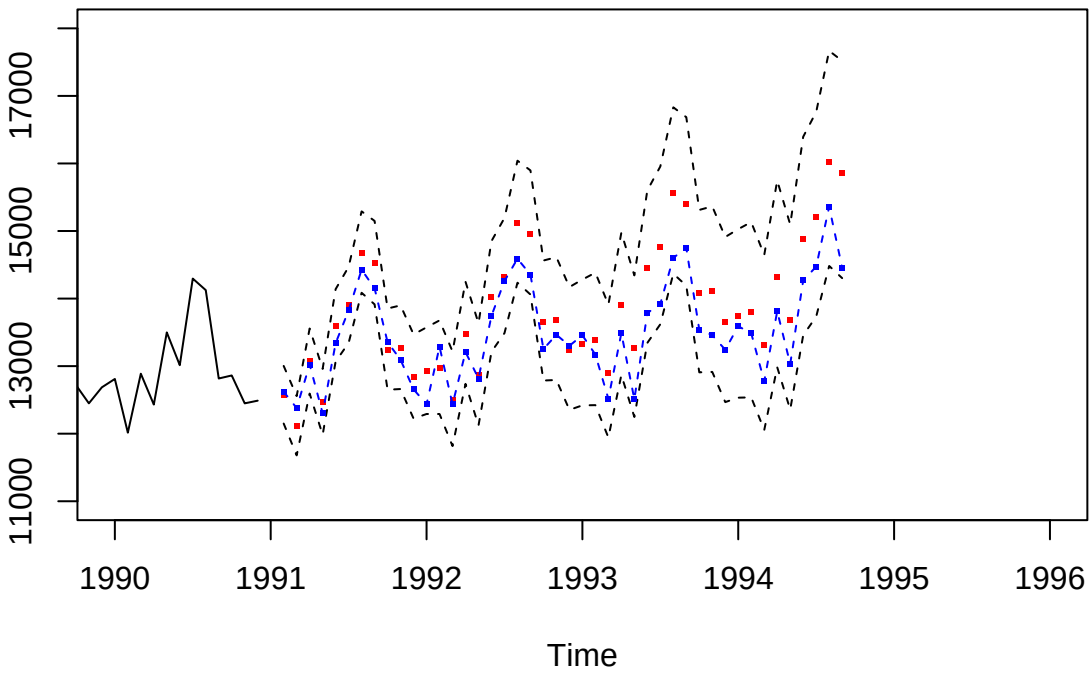
As this model has no AR part, this model is stationary. Then for invertible, apparently the roots of the MA part $(1-0.7363B^{12})(1-0.6714B) = 0$ are all outside the unit circle (Becuase the MA1 and SMA1 parameter absolute value both smaller than 1), so this model is invertible. So I think this model is good enough for forecasting.

Forecast 1991-1995 productions

The forecast of 1991 to 1995 productions:



It is not displayed very well, so I just show the plot with a smaller and proper range from 1992:



From the forecast result, all of the real data points colored in blue are inside the 95% confidence interval. And many predicted red points are close to the true value. That indicates my model forecast is doing a great job. For the future trend, the seasonal change is preserved, and there is a steady increase in mean for the next three year.

Conclusion

In this project, I applied the models and tests from PSTAT 174 class on the data set in the tsdl library. In the first part, I found that Box-Cox transformation is required and I need additional lag 1 and lag 12 difference. Then I inspected the sample ACF, PACF plot and propose two candidate models. After calculating the AICc, Our model is: (After box cox transformation of lambda 0.384‘)

$$X_n - X_{n-1} - X_{n-12} + X_{n-13} = Z_n - 0.6714Z_{n-1} - 0.7363Z_{n-12} + 0.4944Z_{n-13}, Z_n \overset{iid}{\sim} N[0, 0.4201]$$

Then I checked the residual of the model, the sample ACF, PACF, some independence tests, normality tests and proved that the residual is quite alike white noise, and check the roots. All of the tests were valid so I could use this model to forecast. Last, In the forecast part, all of the true observations were in the 95% confidence interval for the forecast. For the future trend, the seasonal change was preserved, and there is a steady increase in mean for the next three year.

References

Introduction to the tsdl package, https://pkg.yangzhuoranyang.com/tsdl/articles/tsdl.html

Appendix: Rcode

```
knitr::opts_chunk$set(echo = FALSE,message = FALSE,warning = FALSE
)
library(forecast)
library(MASS)
library(tseries)
library(car)
library(dplyr)
library(ggplot2)
library(tsdl)
data = tsdl[[122]]
attributes(tsdl[[122]])
train = ts(data[(12*12+1):(12*36)],frequency = 12,start = c(1967,1))
test = data[-(1:(12*36))]
autoplot(train,ylab = "million kilowatt hours") + ggtitle("Monthly electricity production in Australia")
#Box Cox
lambda = boxcox(train~1)
optlambda = lambda$x[which.max(lambda$y)]
train1 = (train^optlambda - 1)/optlambda
autoplot(train1) + ggtitle("transformaed data")
library(dplyr)
train2 = train1 %>% diff(lag=12)  %>% diff(lag=1)
autoplot(train2)
adf.test(train2)
# sample ACF PACF plots
acf(train2, main = "Autocorrelation",lag.max = 48)
pacf(train2, main = "Partial Autocorrelation",lag.max = 48)
# model selection
AICc = 3000
arimalist = list(c(0,1,0),c(1,1,0),c(2,1,0),
   c(3,1,0),c(0,1,1),c(1,1,1),c(2,1,1),c(3,1,1))
seasonallist = list(c(0,1,0),c(1,1,0),c(2,1,0),c(0,1,1),c(1,1,1),c(2,1,1),c(0,1,2),c(1,1,2),c(2,1,2))
AICcs = matrix(NA,length(arimalist),length(seasonallist))
bestmodel = 1
for(i in 1:length(arimalist)){
  for(j in 1:length(seasonallist)){
    result = tryCatch({
      fit = Arima(train1,order=arimalist[[i]],
       seasonal = seasonallist[[j]],
       method = "CSS-ML")
      AICcs[i,j] = fit$aicc

      if(fit$aicc < AICc){
      AICc = fit$aicc
      bestmodel = fit
    }
  }, error = function(e) {
      fit = list()
      fit$aicc = NA
  })
  }
}
```

```
rownames(AICcs) = c("ARIMA(0,1,0)","ARIMA(1,1,0)","ARIMA(2,1,0)",
 "ARIMA(3,1,0)","ARIMA(0,1,1)","ARIMA(1,1,1)",
 "ARIMA(2,1,1)","ARIMA(3,1,1)")
colnames(AICcs) = c("(0,1,0)_12","(1,1,0)_12","(2,1,0)_12",
                    "(0,1,1)_12","(1,1,1)_12","(2,1,1)_12",
                    "(0,1,2)_12","(1,1,2)_12","(2,1,2)_12")
knitr::kable(AICcs, caption = "AICcs of the candidate models")
fit1 = bestmodel
summary(fit1)
fit2 =Arima(train1,order=c(0,1,1),
       seasonal = c(1,1,1),
       method = "CSS-ML")
summary(fit2)
res = resid(fit1)
ggtsdisplay(res)
# check normality
hist(res)
qqnorm(res)
qqline(res)
shapiro.test(res)
# independence
pv = c()
pv[1]=Box.test(res, lag = 17, type = c("Box-Pierce"),fitdf = 2)$p.value
pv[2]=Box.test(res, lag = 17, type = c("Ljung-Box"),fitdf = 2)$p.value
pv[3]=Box.test(res^2, lag = 17, type = c("Ljung-Box"),fitdf = 0)$p.value
df = data.frame(tests = c("Box-Pierce","Ljung-Box","McLeod-Li"), p.value = pv, df = c(15,15,17))
knitr::kable(df)
# Forecast
pred = predict(fit1, n.ahead=44)$pred
se = predict(fit1, n.ahead=44)$se
pred1 = exp(log(pred * optlambda + 1)/optlambda)
pred1l = exp(log((pred-1.96*se) * optlambda + 1)/optlambda)
pred1u = exp(log((pred+1.96*se) * optlambda + 1)/optlambda)
ts.plot(train, xlim = c(1969,1996))
points(1991+(1:44)/12,pred1,col = "red")
lines(1991+(1:44)/12,pred1l,lty=2)
lines(1991+(1:44)/12,pred1u,lty=2)
points(1991+(1:44)/12,test,col = "blue")
lines(1991+(1:44)/12,test,col = "blue",lty=2)
ts.plot(train, xlim = c(1990,1996), ylim =c(11000,18000),title = "Forecast 1994-1997 productions")
points(1991+(1:44)/12,pred1,col = "red",pch  = 46,cex = 3)
lines(1991+(1:44)/12,pred1l,lty=2)
lines(1991+(1:44)/12,pred1u,lty=2)
points(1991+(1:44)/12,test,col = "blue",pch  = 46,cex = 3)
lines(1991+(1:44)/12,test,col = "blue",lty=2)
```