Analysis and Prediction of Electricity Production based on Past 33 years in India

Suryasis Jana (Roll no - 201447)
Sayan Bhowmik (Roll no - 201409)
Arijit Ghosh (Roll no - 201276)
Soumyadip Sarkar (Roll no - 201431)

November 16, 2021

Contents

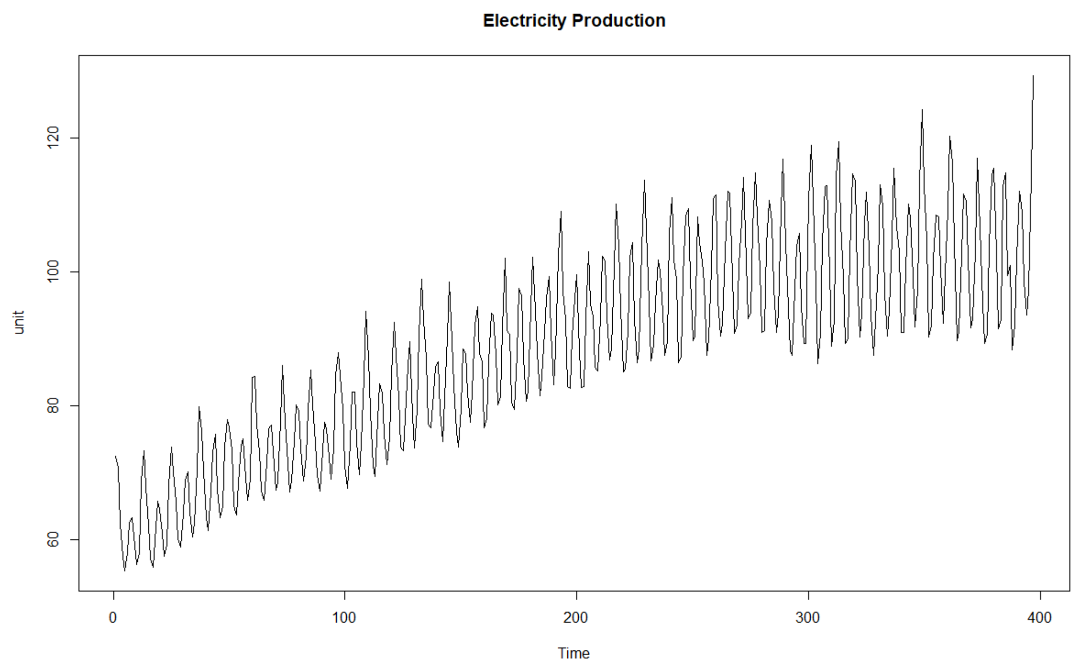
1 Introduction

A time series is basically series of observations recorded sequentially over a period of time i.e. it is a collection of observations indexed in time order, represented as $(t, Y_t);\ t = 1, 2, ..., n$. Time Series analysis accounts for the fact that the data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be explored and used to forecast future observation.

2 Data Description

We have taken the following data on monthly production of electricity in India. There are 397 datapoints from January 1985 to January 2018. Source of this dataset is `https://www.kaggle.com/shenba/time-series-datasets`

3 Plot of the raw data



4 Objectives

1. Testing the presence of trend.
2. Testing the presence of seasonality.
3. Determination the period of seasonality.
4. If trend and seasonality exist, then estimation and elimination of them.
5. Checking the randomness of the deseasonalised and detrended data.
6. Finding out the probability model that well fits the random data and estimation of its parameters.
7. Forecasting.

5 Testing the presence of trend (Using Relative Ordering Test)

This is a **non-parametric test** used to check the presence of trend component in a time series. If we denote the null hypothesis by H_0 and alternate hypothesis by H_1, then

$$H_0 : \text{ Trend is absent} \qquad \text{vs} \qquad H_1 : \text{ Trend is present}$$

Let Q be the number of points of discordance which means Q is count of those pairs (Y_i, Y_j) for which either $Y_i > Y_j$ for $i < j$ or $Y_i < Y_j$ for $i > j$. Let us define T as,

$$T = 1 - \frac{4Q}{n(n-1)}$$

Then our test statistic becomes,

$$Z = \frac{T}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}$$

Asymptotic test for H_0: no trend is based on the statistic Z.Under H_0 Z asymptotically follows N(0,1). Therefore we reject the null hypothesis at level of significance α if

$$|Z_{obs}| \;>\; \tau_{\frac{\alpha}{2}}$$

Here we have $|Z_{obs}| \;=\; 19.1257$ and $\tau_{0.025} \;=\; 1.96$

$$|Z_{obs}| \;>\; \tau_{0.025}$$

So we reject null hypothesis at 5% level of significance and hence conclude that trend is present.

6 Testing for presence of seasonality (using Friedman's test)

H_0 : no seasonality present vs H_1 : Seasonality is present

Once we estimate the trend, we eliminate it from the process and assign ranks for the months in each year. Here we define

$$M_{ij} : \text{ the rank corresponding to ith month of jth year}$$

$$M_i := \sum_{j=1}^{c} M_{ij} \qquad i = 1(1)12, \ j = 1(1)c, \ c = 33$$

The test statistic for our test is given by,

$$X = 12\sum_{i=1}^{12} \frac{(M_i - \frac{c(12+1)}{2})^2}{c \times 12 \times 13}$$

Under H_0 the statistic asymptotically follows χ^2_{11} distribution. If we have $X_{obs} > \chi^2_{\alpha,11}$ we reject null hypothesis at level of significance α.

In this case we have got $X_{obs} = 2882.333$ and $\chi^2_{0.05,11} = 19.68$
So we reject the null hypothesis at 5% level and conclude that seasonality is present in the process we have in the dataset.

7 Determination of period of seasonality

Here firstly, we have found out the process $\nabla_{12}X_t$ and then shown that the process we obtained forms a random process using turning point test.

Consider the $\nabla_{12}X_t$ process.We want to test,

H_0 : Process is random vs H_1 : Process is not random

Let P denotes the total numbers of turning points in the series $\nabla_{12}X_t$

Then our test statistic is given by,

$$Z = \frac{P - \frac{2}{3}(n-2)}{\sqrt{\frac{16n-29}{90}}}$$

Under H_0 Z follows N(0,1) asymptotically and reject null hypothesis if $|obs\ Z| > \tau_{\alpha/2}$
Here we get mod value of observed Z = 0.70 and $\tau_{0.025} = 1.96$

So we conclude that the null hypothesis is accepted here at 5% level of significance and the period of seasonality is 12.

8 Estimation and Elimination of components of time series (Using fast trend method)

8.1 Step-1

Here we first obtain the rough estimates of the trend component with a 12 point moving average.
Let Y_t be our original data where t = 1, 2, ..., 396 [Here we neglect the data for January 2018 for the sake of calculation] So d = 12 here.
The 12 point moving average estimates are given by,

$$\hat{m_t} = \frac{1}{12}(\frac{1}{2}Y_{t-6} + Y_{t-5} + \cdots + Y_t + \cdots + Y_{t+5} + \frac{1}{2}Y_{t+6}) \qquad \text{where } 6 < t < 391$$

And the first and the last 6 time points or observations are estimated using endpoint padding method.

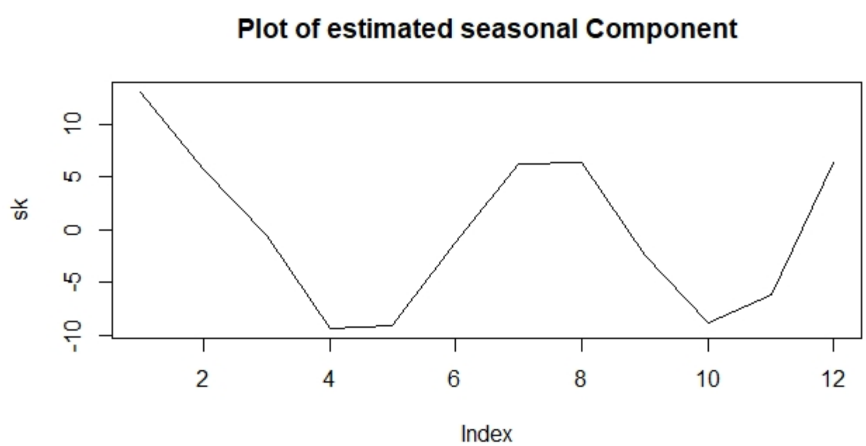
8.2 Step -2

Now we proceed to estimate the seasonal component.
For each k = 1,2,...,12 we define,

$$W_k = \frac{1}{J}\sum_{j=1}^{J}(y_{k+(j-1)d} - \hat{m}_{k+(j-1)d})$$

Here J = 33. Then the final estimates of seasonal component are given by,

$$\hat{s}_k = W_k - \frac{1}{d}\sum_{k=1}^{d} W_k \qquad k = 1, 2, \cdots, 12$$



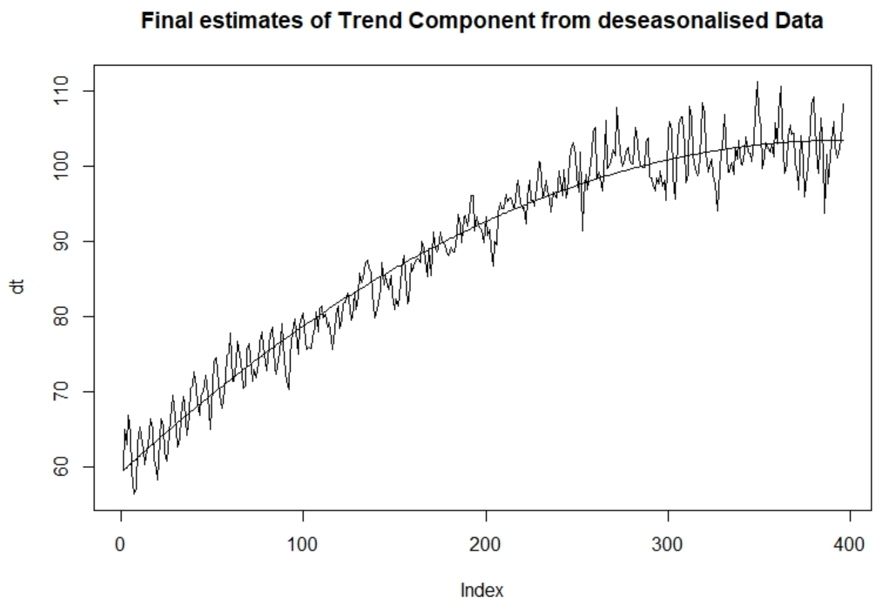
8.3 Step-3

Next we calculated deseasonalised data as

$$d_t \;=\; Y_t - \hat{S}_t \qquad\qquad t = 1, 2, \cdots, 396$$

8.4 Step-4

Next we re-estimated the trend component for $\{d_t\}_{t=1}^{396}$using a quadratic polynomial trend.



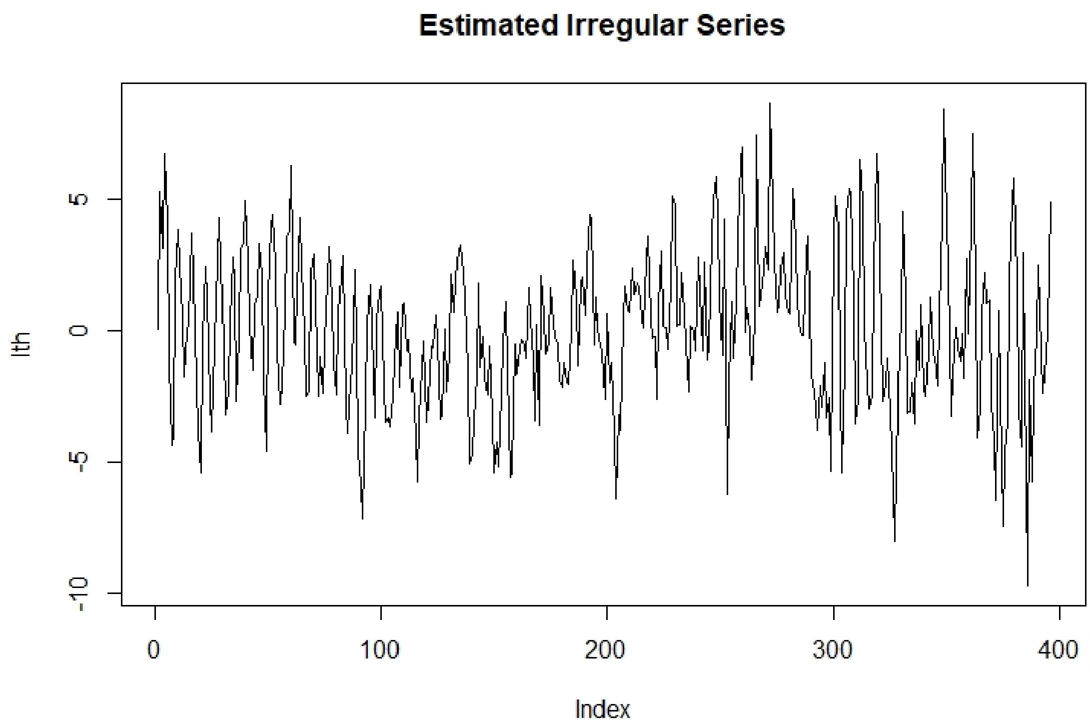
The estimated trend for the d_t series is given by,

$$\hat{\hat{m}}_t = 59.1781 + 0.2235t - 0.000282t^2 \qquad t = 1, 2, \cdots, 396$$

8.5 Step-5

Next our estimated irregular series is given by,

$$\hat{I}_t = Y_t - \hat{\hat{m}}_t - \hat{S}_t \qquad\qquad t = 1, 2, \cdots, 396$$



Next we applied Turning Point Test on $\hat{I}_t$ and found that $|obs\ Z| = 1.01$ and $\tau_{.025} = 1.96$

Therefore we accept the null hypothesis about the randomness of the series $\hat{I}_t$ at 5% level of significance

9 Analysis of the Estimated Irregular Series $\{\hat{I}_t\}$

We now check whether the series $\{\hat{I}_t\}_{t=1}^{396}$ is from a WN process or not. So the null hypothesis and the alternate hypothesis can be written as,

H_0 : $\hat{I}_t$ is from a WN process against H_1 : H_0 is false

The test statistic here is

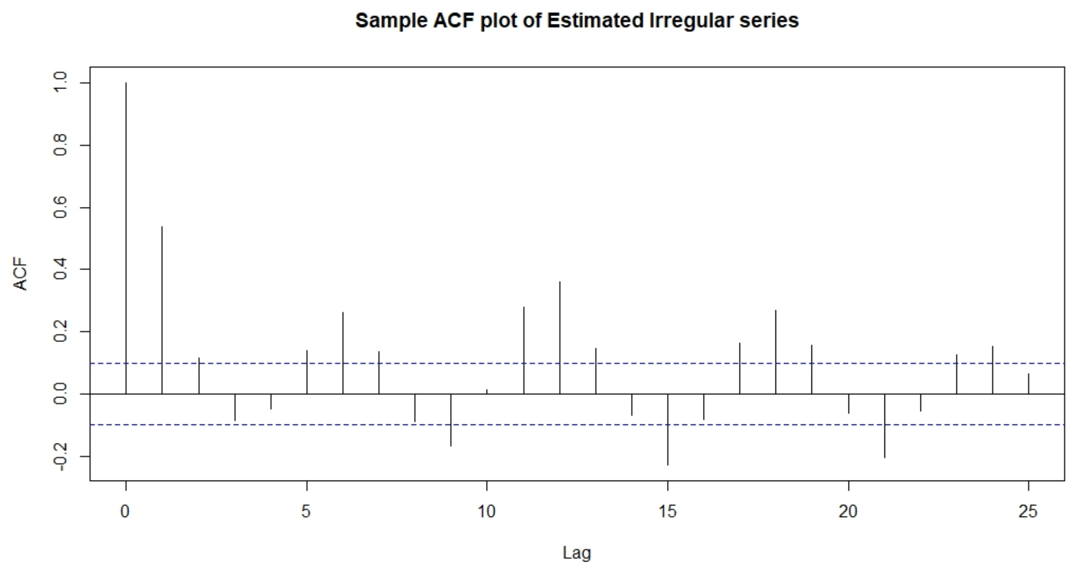
$$Z = \sqrt{n}\hat{\rho}_1$$

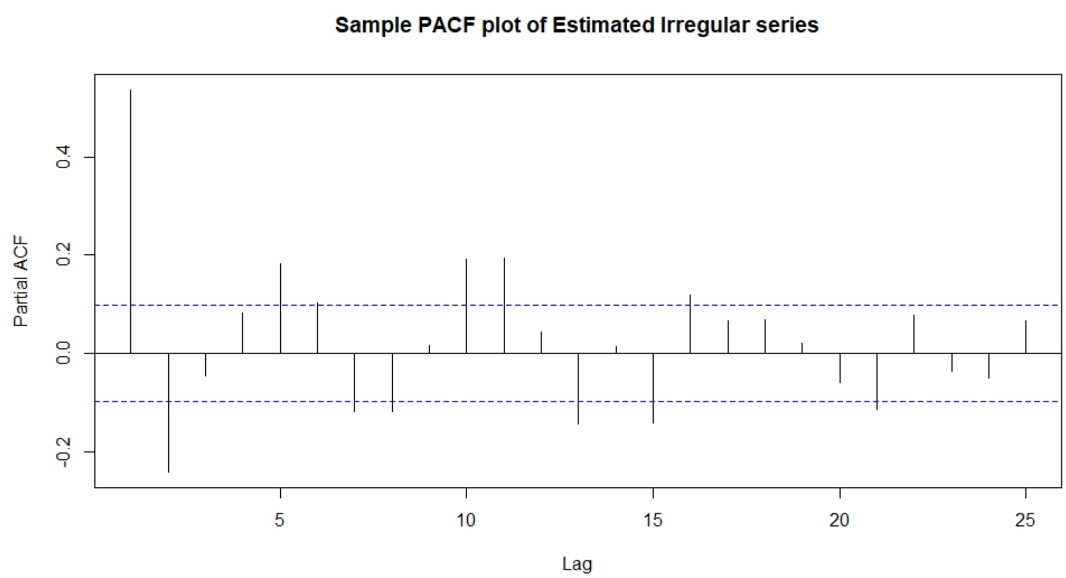
Under H_0, the statistic follows N(0,1) asymptotically. We reject the null hypothesis at level α iff

$$|Z_{obs}| > \tau_{\alpha/2}$$

We have got $|Z_{obs}| = 10.72$ and $\tau_{0.025} = 1.96$. Therefore, at 5% level of significance, we may conclude that $\hat{I}_t$ is not from a WN process.

Now we have obtained the ACF and PACF plots of the estimated series $\hat{I}_t$





From both the ACF and PACF plots we note that both the plots 'tail off' i.e. the spike never cuts off after a certain lag. So we suspect that the series follows an **ARMA(p,q)** process where both p and q are greater than 0.

9.1 Estimation of order of ARMA model

To find out the appropriate order of ARMA(p,q) we shall calculate the value of **Akaike Information Criterion(AIC)** for various values of p and q and shall choose the value for which **AIC is minimum**.
The formula for AIC of an ARMA(p,q) model is given by

$$AIC(p,q) = -2log\hat{L} + 2(p+q+1)$$

```
> round(aic,2)
```

p \ q	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1990.61	1854.76	1837.39	1838.82	1838.73	1834.85	1836.72	1827.35	1824.30	1826.03	1826.38	1826.28	1804.08	1804.21	1800.63
1	1857.24	1840.53	1839.12	1840.43	1839.72	1836.83	1831.13	1826.95	1826.19	1827.34	1828.22	1820.89	1803.63	1805.62	1802.43
2	1834.29	1835.88	1834.24	1794.28	1784.17	1771.38	1764.39	1766.33	1768.28	1770.28	1772.28	1773.79	1775.46	1804.50	1765.71
3	1835.56	1837.25	1766.19	1768.05	1789.84	1784.98	1764.21	1763.40	1764.70	1765.35	1766.65	1767.51	1769.26	1770.76	1766.62
4	1835.09	1829.14	1768.06	1768.08	1770.07	1771.43	1763.22	1764.59	1765.52	1767.16	1765.91	1769.26	1770.66	1771.90	1771.53
5	1823.27	1823.51	1769.92	1770.07	1768.51	1759.15	1772.87	1774.28	1774.87	1768.67	1767.87	1763.29	1771.08	1773.70	1770.83
6	1820.66	1820.60	1767.68	1771.30	1766.75	1770.75	1774.81	1762.54	1761.71	1765.32	1763.06	1769.87	1755.51	1757.11	1757.19
7	1816.45	1816.05	1768.72	1765.21	1767.21	1775.28	1761.67	1763.52	1764.97	1758.80	1765.06	1772.19	1757.05	1759.07	1756.23
8	1812.42	1814.41	1768.04	1767.21	1769.22	1770.02	1763.49	1775.33	1776.01	1763.55	1765.33	1771.25	1756.33	1759.74	1758.22
9	1814.39	1809.02	1768.24	1768.54	1768.80	1767.09	1765.07	1773.50	1775.24	1766.53	1762.92	1760.76	1752.50	1754.39	1756.55
10	1799.60	1792.51	1798.29	1765.36	1765.58	1764.02	1758.52	1767.88	1768.21	1769.03	1764.56	1759.90	1752.10	1753.54	1755.70
11	1787.50	1789.04	1769.53	1770.70	1761.25	1759.12	1759.58	1761.43	1763.38	1764.76	1754.55	1751.33	1752.32	1754.20	1753.37
12	1788.31	1777.88	1777.80	1781.09	1765.79	1754.04	1756.00	1754.97	1758.66	1750.74	1751.03	1751.05	1752.88	1755.35	1755.01
13	1781.55	1777.57	1777.85	1768.48	1767.74	1756.42	1756.44	1757.46	1760.48	1760.89	1751.19	1752.86	1744.51	1745.71	1747.55
14	1783.54	1779.17	1779.65	1766.18	1764.17	1757.70	1757.17	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

From the above table we note that AIC is minimum for **p = 13 and q = 12**
So we conclude that estimated irregular series is from a **ARMA(13,12) process**.

9.2 Estimation of Model Parameters

Since the minimum value of AIC for the ARMA model is 1744.51 for order p=13 and q=12, our model is ARMA(13,12). Therefore we have to estimate 13 AR parameters, 12 MA parameters, the intercept term and the error variance by the **method of Maximum Likelihood**.

The estimates of the parameters obtained are given below in the following table:

```
> model

Call:
arima(x = It, order = c(p, 0, q), method = "ML")

Coefficients:
          ar1     ar2      ar3      ar4     ar5     ar6     ar7     ar8
      -0.0627  0.3720  -0.4610  -0.2932  0.1624  0.0568  0.0639  0.4145
s.e.   0.4413  0.2877   0.0845   0.2494  0.1340  0.0941  0.1100  0.0881
          ar9     ar10    ar11    ar12     ar13     ma1      ma2     ma3
      -0.0302  -0.1529  0.6665  0.3734  -0.3808  0.6340  -0.0511  0.4701
s.e.   0.2402   0.1723  0.0699  0.3052   0.1270  0.4513   0.0923  0.0780
         ma4     ma5     ma6     ma7      ma8      ma9    ma10     ma11
      0.6130  0.1587  0.1553  0.1133  -0.5817  -0.3384  0.1771  -0.6187
s.e.  0.3087  0.0970  0.1160  0.0993   0.1122   0.3025  0.1023   0.0916
         ma12  intercept
      -0.6667     0.0501
s.e.   0.3716     0.3802

sigma^2 estimated as 4.017:  log likelihood = -845.25,  aic = 1744.51
```

9.3 Estimated Model

Replacing the above mentioned parameters our model looks like,

$$\begin{aligned}
\mathbf{Y_t} &= \mathbf{0.0501 - 0.0627Y_{t-1} + 0.3720Y_{t-2} - 0.4610Y_{t-3} - 0.2932Y_{t-4} + 0.1624Y_{t-5}} \\
&\mathbf{+0.0568Y_{t-6} + 0.0639Y_{t-7} + 0.4145Y_{t-8} - 0.0302Y_{t-9} - 0.1529Y_{t-10} + 0.9995Y_{t-11}} \\
&\mathbf{+0.3734Y_{t-12} - 0.3808Y_{t-13} + \epsilon_t + 0.6340\epsilon_{t-1} - 0.0511\epsilon_{t-2} + 0.4701\epsilon_{t-3}} \\
&\mathbf{+0.6130\epsilon_{t-4} + 0.1587\epsilon_{t-5} + 0.1553\epsilon_{t-6} + 0.1133\epsilon_{t-7} + 0.5817\epsilon_{t-8} - 0.3384\epsilon_{t-9}} \\
&\mathbf{+0.1771\epsilon_{t-10} - 0.6187\epsilon_{t-11} - 0.6667\epsilon_{t-12}}
\end{aligned}$$

10 Forecasting

One of the primary objective of building a model for a time series is to be able to forecast the values for that series at future time points. We also want to assess the precision of the forecast. For the most of the time we shall assume that the model is known exactly, including the specific values for all parameters. Although this is never true in practice, for large sample size, the use of estimated parameters doesn't seriously affect forecast.

We use the **Best Linear Predictor (BLP)** approach here. Given n data points $X_1, X_2, \cdots, X_n$, we want to predict X_{n+h}. The BLP function is given by ,

$$P_n(X_{n+h}) = a_0 + \sum_{i=1}^{n} a_i X_{n+1-i}$$

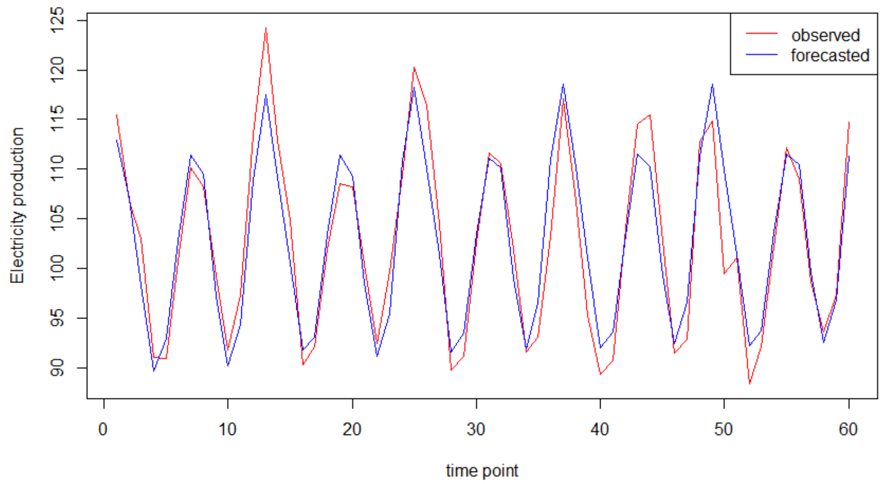
where a_i's are such that $E(X_{n+h} - P_n(X_{n+h}))^2$ is minimum among all such linear combinations.

Using this forecasting technique we have predicted the values at some future points. Below is the forecasted values for next 96 months i.e., 8 years:

	Month	Forecasted Values		Month	Forecasted Values		Month	Forecasted Values		Month	Forecasted Values
1	Jan-18	119.6905	25	Jan-20	117.7468	49	Jan-22	118.1407	73	Jan-24	116.5461
2	Feb-18	111.3616	26	Feb-20	109.4145	50	Feb-22	109.3602	74	Feb-24	108.2888
3	Mar-18	103.3211	27	Mar-20	101.7051	51	Mar-22	101.1867	75	Mar-24	100.5213
4	Apr-18	91.16328	28	Apr-20	91.71138	52	Apr-22	91.41175	76	Apr-24	91.22876
5	May-18	94.02892	29	May-20	93.56619	53	May-22	92.53979	77	May-24	92.21117
6	Jun-18	101.9815	30	Jun-20	103.8362	54	Jun-22	102.2086	78	Jun-24	101.5353
7	Jul-18	111.8443	31	Jul-20	111.2727	55	Jul-22	110.3585	79	Jul-24	109.6221
8	Aug-18	110.11	32	Aug-20	110.3172	56	Aug-22	109.6456	80	Aug-24	108.6088
9	Sep-18	99.29553	33	Sep-20	99.48411	57	Sep-22	99.57542	81	Sep-24	98.09773
10	Oct-18	92.98269	34	Oct-20	91.94335	58	Oct-22	92.0956	82	Oct-24	90.89895
11	Nov-18	97.20849	35	Nov-20	96.0182	59	Nov-22	95.79209	83	Nov-24	94.62046
12	Dec-18	110.9592	36	Dec-20	109.9491	60	Dec-22	109.9547	84	Dec-24	108.8723
13	Jan-19	118.3788	37	Jan-21	118.0499	61	Jan-23	117.0428	85	Jan-25	116.2075
14	Feb-19	110.0729	38	Feb-21	110.008	62	Feb-23	108.6452	86	Feb-25	107.9821
15	Mar-19	101.0413	39	Mar-21	101.6663	63	Mar-23	100.5865	87	Mar-25	100.2341
16	Apr-19	91.73076	40	Apr-21	91.99455	64	Apr-23	91.09561	88	Apr-25	90.49451
17	May-19	92.62818	41	May-21	93.47131	65	May-23	92.588	89	May-25	91.36461
18	Jun-19	103.1972	42	Jun-21	102.5873	66	Jun-23	102.0579	90	Jun-25	100.6909
19	Jul-19	112.2537	43	Jul-21	110.7121	67	Jul-23	110.2774	91	Jul-25	108.7323
20	Aug-19	110.1297	44	Aug-21	109.581	68	Aug-23	109.5947	92	Aug-25	108.0507
21	Sep-19	100.4182	45	Sep-21	99.13229	69	Sep-23	98.86871	93	Sep-25	97.72463
22	Oct-19	92.71519	46	Oct-21	92.32388	70	Oct-23	91.452	94	Oct-25	90.64823
23	Nov-19	96.3459	47	Nov-21	95.91503	71	Nov-23	95.08277	95	Nov-25	94.42751
24	Dec-19	110.2846	48	Dec-21	110.405	72	Dec-23	109.0496	96	Dec-25	108.2961

11 Model accuracy

To check whether our model is good enough or not for the sake of prediction or forecasting, we cross checked the model for our data dividing it into 2 parts - train set and test set. The train set contains the first 336 data points and the test set contains the rest 60 data points (we ignored 397th data point for the sake of simplicity in calculation). On repeating the same process on the train set, the model having minimum AIC value turns out to be ARMA(13,12). Now when we imposed the model on the test set the plots of observed and predicted values look like the graphs given below.



And we also calculated the MSE (Mean Squared Error) and MAPE (Mean Absolute Percent Error) for our model on the test set. They turn out to be 10.51341 and 2.502531 respectively, which shows that our model is good enough to predict the electricity production in future.

12 R-code

```
# Sourse: https://www.kaggle.com/shenba/time-series-datasets?select=
Electric_Production.csv
data=read.csv("Electric_Production.csv")
colnames(data)=c("date","unit")
attach(data)
plot(unit,type="l",xlab="Time",main="Electricity Production")

##  A) Testing for the presence of trend using Relative ordering test:
n=length(unit)
count=0
for(i in 1:(n-1)){
for(j in (i+1):n){
if(unit[i]>unit[j]){
count=count+1
}}}
count
tau1=1-((4*count)/(n*(n-1)))
tau1
var_tau1=(2*(2*n+5)/(9*n*(n-1)))
z=tau1/sqrt(var_tau1)
z
qnorm(1-0.025)
#  Reject H0 and conclude that trend is present

## B) Test for present of Seasonality Friedman Test

Yt=unit
t=1:length(Yt)
Xt=Yt-predict(lm(Yt~t))
plot(Xt,type="l",xlab="Time",main="Electricity Consumption")
Xt=Xt[-397]
length(Xt)
A=matrix(Xt,nrow=12,byrow=F)
Mi=apply(A,1,sum)
c=33
r=12
X=12/(c*r*(r+1))*sum((Mi-c*(r+1)/2)^2)
X
qchisq(1-0.05,r-1)
# H0 is rejected and conclude that seasonality is present

## C) Determination of Period of Seasonality
D12.Xt=Xt[13:396]-Xt[1:384]
plot(D12.Xt,type="l",xlab="Time",main="Electricity Consumption")
```

```
u=0
rpm=D12.Xt
for(i in 2:(384-1))
{
if(((rpm[i]>rpm[i+1])&&(rpm[i]>rpm[i-1]))||((rpm[i]<rpm[i+1])
&&(rpm[i]<rpm[i-1])))
{
u=u+1
}
}
u
e_u=2*(n-2)/3
v_u=(16*n-29)/90
z=(u-e_u)/v_u
z
qnorm(1-0.025)

# Since D12 makes he series purely random so the period of seasonalaty is 12

### So, Let us move to estimation of components of time series

#Rough Estimation of trend using 12 point moving average
Yt=Yt[-397]
Y1t=array(0)
for(i in (1:385)){
Y1t[i]=mean(Yt[i:(i+11)])
}

j=1
a1=array(0)
for(i in 7:2){
a1[j]=(i*Yt[1]+sum(Yt[2:(12-(i-1))]))/12
j=j+1
}
length(a1)

j=1
a2=array(0)
for(i in 2:7){
a2[j]=(i*Yt[396]+sum(Yt[395:(385+(i-1))]))/12
j=j+1
}
length(a2)

Y11t=c(a1,Y1t,a2)
```

```
length(Y11t)

m1t=array(0)
for(i in (1:396)){
m1t[i]=mean(Y11t[i:(i+1)])
}

length(m1t)   # m1t is the moving average values, which are
the rough estimates of trend

D1t=(Yt-m1t)
j=1:33
wk=array(0)
for(k in 1:12){
wk[k]=mean(D1t[k+(j-1)*12])
}
length(wk)
sk=wk-mean(wk)  # This is the estimate of seasonal component

st=rep(sk,times=33)
plot(st,type="l")

D2t=Yt-st # Final deseasonalized data
plot(D2t,type="l",ylab = "dt",main="Final estimates of Trend Component from
deseasonalised Data")

t=1:396
t1=t^2
m2t=predict(lm(D2t~t+t1))  # This is the estimate of Trend component
lines(t,m2t)


It=Yt-m2t-st
plot(It,type="l",ylab = "Ith", main = "Estimated Irregular Series")

## Testing for irregular component
n=396
u=0
for(i in 2:395)
{
  if(((It[i]>It[i+1])&&(It[i]>It[i-1]))||((It[i]<It[i+1])&&(It[i]<It[i-1])))
  {
    u=u+1
  }
}
u
```

```
e_u=2*(n-2)/3
v_u=(16*n-29)/90
z=(u-e_u)/v_u
z
qnorm(1-0.025)

# So it is accepted that It is a stationary process


### Checking whether It is WN or not
length(It)
sqrt(396)*cor(It[1:395],It[2:396])
qnorm(1-0.025)
# Reject H0 and conclude that It series is not WN

acf(It,main="Sample ACF plot of Estimated Irregular series")
pacf(It,main="Sample PACF plot of Estimated Irregular series")


###########
aic=array(0,dim=c(15,15))
for(i in 0:14)
{
  for(j in 0:14)
  {
    aic[(i+1),(j+1)]=0
    model=arima(It,order=c(i,0,j),method = "ML")
    aic[(i+1),(j+1)]=model$aic
  }
}
round(aic,2)
min(aic)

# p=13,q=12
aic=as.numeric(aic)
i=14
for(j in 10:14){
    aic[(i+1),(j+1)]="NaN"
}

aic=matrix(aic,nrow=15,byrow=F)
rownames(aic)=c(0:14)
colnames(aic)=c(0:14)
aic

# Looking at aic matrix we choose p=13,q=12
```

```
p=13
q=12
model=arima(It,order=c(p,0,q),method = "ML")
model

### Forecasting Part

fc=predict(model,n.ahead=96,newxreg = NULL,se.fit = F)
fc  # predicted irregular values
sth = rep(sk,times = 8)
sth # predicted irregular values
mth = function(t){
  59.178120 + 0.223504*t - 0.000282*t^2
}
mth(397:492)
finfore=mth(397:492)+sth+fc
dd=seq(as.Date("2018-1-1"),as.Date("2025-12-1"),by="months")
length(dd)
aa=data.frame(dd,finfore)
colnames(aa)=c("Date", "  Forecasted Values")
View(aa)
write.csv(aa,"Forecasted.csv")
```

13 References

1. Class notes of Dr. Amit Mitra, Professor, Department of Mathematics and Statistics, IIT Kanpur.
2. Introduction to Time Series and Forecasting – P. J. Brockwell and R. A. Davies

14 Acknowledgement

Project is like a bridge between the theoretical learning and practical working. With this will, we all started this project.First of all, We are feeling oblige in taking the opportunity to express our sincere thanks and gratitude to our professor **Dr. Amit Mitra** who gave us the golden opportunity to execute this Time Series Project and also helped us to complete it. We also feel delighted to thank all our seniors and classmates who are also somehow responsible for the successful completion of this project.