Monthly US Electricity Production

Rayaleswar Mondrety and Chris Nottke

Abstract

Electricity has been an important part of human history. Most of us use electricity everyday, in order to go on with our daily lives, and the United States electricity production plants produce electricity to meet this consumer demand. Monthly electricity production data from the US Energy Information Administration was first transformed with a logarithmic transformation, and then the seasonal difference at lag 12 was taken. This transformed data was modeled with ARIMA(2,0,1)(0,0,1)[12], and diagnostics were computed to determine the goodness of fit of the model to the data. The data and model were then used to forecast values 5 years after the end of the data. This type of analysis is important, due to the ever increasing reliance on electricity.

Content

Introduction:

Humans have known about electricity for over two millennia, in primitive forms like static on amber stone and later some magicians took advantage of this static phenomenon and developed parlor tricks for entertainment in the early 16th century. Benjamin Franklin is credited with the discovery of electricity, because of his famous kite experiment[1](Energy, 2022). But the electricity as we know and use it, at least in our opinion, can be given to first Nikola Tesla and Thomas Edison. Their famous Alternating Current vs Direct Current race to electrification of the USA to showcase the superiority of their technologies has been the catalyst to the electrification of the entire world [2] (Lantero, 2014). Electricity as a source of energy is very efficient. If a 1 Kwh of energy is spent we can at least expect about 0.95 Kwh of actual work done when compared to say Gas only 40 - 45% energy in a gallon of gas is converted into useful work.

Electricity is generally generated from sources like Coal, Natural Gas, Nuclear plants, Hydroelectric plants, Solar Farms and some other renewable sources as well. Life without electricity is unimaginable in the modern day, and with rapid increase in the rate of industrialization, urbanization, agriculture and transportation the consumption of electricity is increasing rapidly. In houses across the country the number of air conditioners are increasing due to Global warming and causing an increase in electricity consumption. As the consumption of electricity increases the production of electricity also increases[3](U.S. Energy Information Administration - EIA - Independent Statistics and Analysis, n.d.).

So, we wanted to check the rate at which the electricity production has increased from 1970's to 2005 and perform some time series analysis on the data and do a forecast. The dataset is taken from the TSA R-package that was originally from the U.S. Energy Information Administration data.

This dataset has some seasonality and we have wanted to answer the following questions:

1. Can we fit a Seasonal ARIMA Model
2. Forecast the Electricity production for 5 years.

Data Visualization

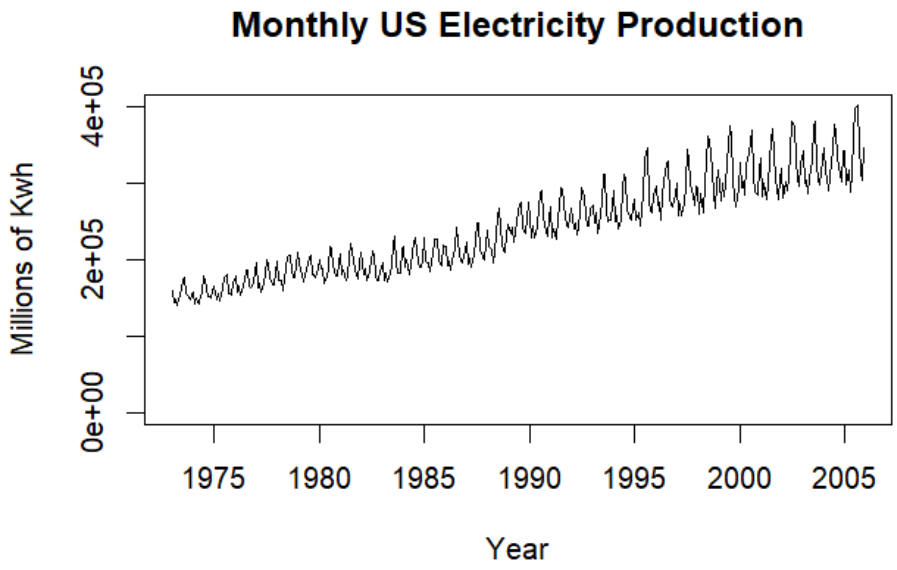

Figure 1: Time series data of monthly US electricity production from all sources, from January 1973 to December 2005.

Electricity production in kilowatt hours from all sources was captured each month to compile the data. There were fortunately no missing values, and no outliers were present. The data exhibits an upward deterministic trend with seasonality, which strengthens the case for a transformation. The data has an apparent trend where the variance increases as time increases. The increasing variance also makes it necessary to apply a transformation.

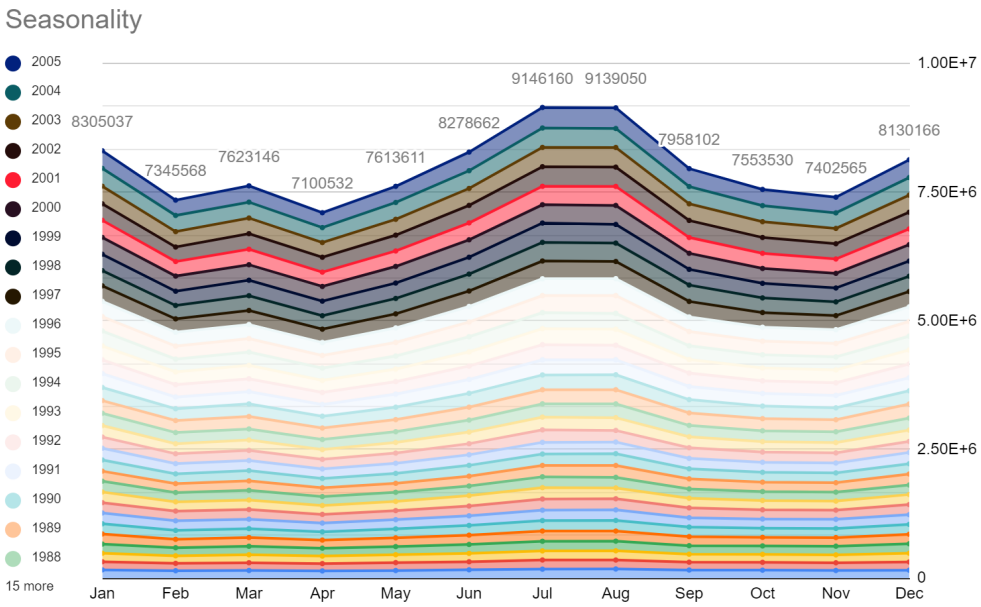

Figure 2: Plot showing seasonality of the data, as well as maximums and minimums in a given year.

It can be seen in the above plot that US monthly production of electricity has a maximum in July and a minimum in April. These descriptive statistics can be explained by weather and how temperate it is. In a hot month, like July, people will be more inclined to use air conditioning units. In a cold month, people will use heating systems. Air conditioning generally uses more electricity than heat, which explains why the maximum is in July, and not in a winter month. The minimum is in April, since this month tends to be temperate.

Transformations

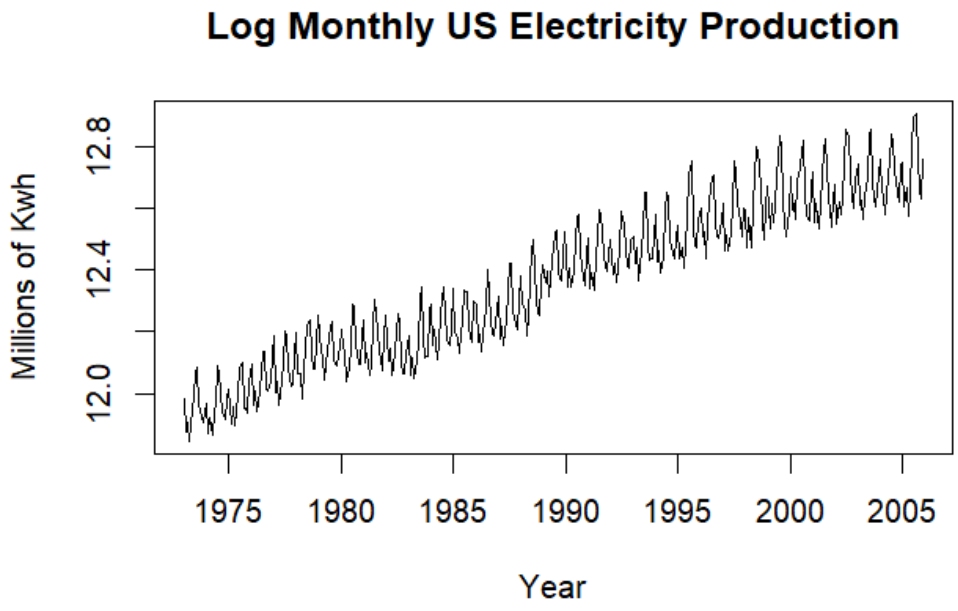

Figure 3: The data underwent a logarithmic transformation in order to stabilize the variance.

Since the data showed increasing variance with time, a logarithmic transformation was applied to the data, which did seem to stabilize the variance throughout the data. The logarithmically transformed data was tested for stationarity using the ADF test, PP test, and the KPSS test. The aforementioned tests all gave the same results of non-stationarity, which begs the need for another transformation, in order to make time series modeling applicable. Next, a seasonal difference had to be taken in order to stabilize the mean, correct seasonality, and to ultimately make the data stationary.

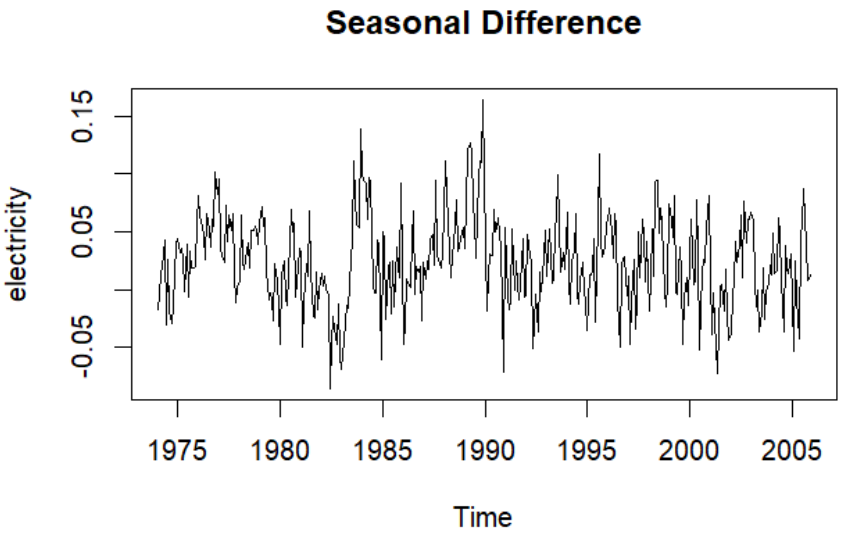

Figure 4: The seasonal difference of the logarithmic transformation of the data shows constant variance and a stabilized mean.

The seasonal difference of the data at the 12th lag was taken in order to account for the seasonality present in the data. The ADF test, PP test, KPSS test all gave the same result of saying the data is stationary. Thus, this transformation was successful in making the data stationary, which makes its candidacy for time series modeling appropriate.

Selecting A Model

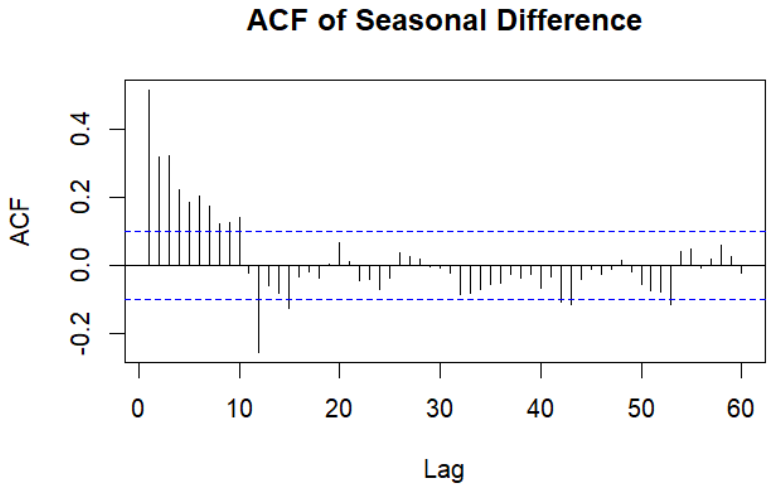

Figure 5: The ACF plot of the transformed data shows a significant autocorrelation at the 12th lag.

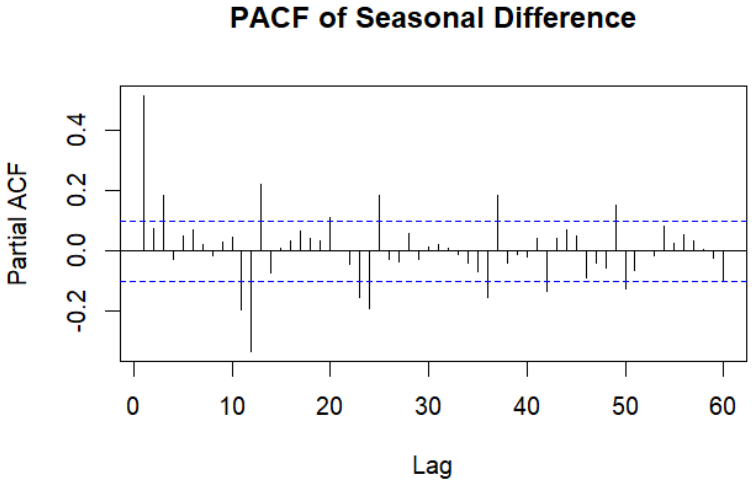

Figure 6: The PACF plot of the transformed data shows autocorrelation very slowly decaying to zero.

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	o	x	o	o
1	x	x	**x**	**o**	o	o	o	o	o	x	o	x	x	o
2	x	x	**o**	o	o	o	o	o	o	o	o	x	x	o
3	x	x	x	x	o	o	o	o	o	o	o	x	x	x
4	x	x	x	x	o	o	o	o	o	o	o	x	x	o
5	x	x	x	x	o	o	o	o	o	o	o	x	o	o
6	x	x	x	o	o	o	o	o	o	o	o	x	o	o
7	x	o	x	o	o	o	o	o	o	o	o	x	o	o

Figure 7: EACF plot with bold candidate models.

In order to select a good time series model for the seasonally differenced time series, the ACF plot, PACF plot, and the EACF plot were all investigated. The ACF plot of the transformed data showed significance at the twelfth lag, which helps coming to the conclusion of the seasonal MA order of the model being chosen as one. The PACF shows a slow, alternating decay to zero as the lag increases. The PACF does not have any isolated significant lags, so what can be told from this plot is that the seasonal AR order of the model is zero. The EACF suggests three models that were investigated. Given all of the plots, and the auto.arima function, five candidate models were chosen:

- ARIMA(2,0,1)(2,0,1)[12] - auto.arima
- ARIMA(2,0,1)(0,0,1)[12] - ACF, PACF
- ARIMA(2,0,2)(0,0,1)[12] - EACF
- ARIMA(1,0,3)(0,0,1)[12] - EACF
- ARIMA(1,0,2)(0,0,1)[12] - EACF

From the candidate models, ARIMA(2,0,1)(0,0,1)[12] was selected, since it had an AIC of -1691.68, a AICc of -1691.46, and a BIC of -1667.98. All of the information criteria for this model were smaller than those of the other candidate models. This model's coefficients were also tested to see if they were significantly different from zero with 95% confidence. The results were that all of the coefficients were significant. With all of the aforementioned information, ARIMA(2,0,1)(0,0,1)[12] was chosen to be the best fit model for the seasonally differenced data.

Model Diagnostics

We have chosen the Seasonal ARIMA(2,0,1)(0,0,1)[12] model. In order to see if the residuals of the chosen model are Stationary, homoscedastic, independent, and significance of Autocorrelation function we have performed tsdiag() operation on the residuals.

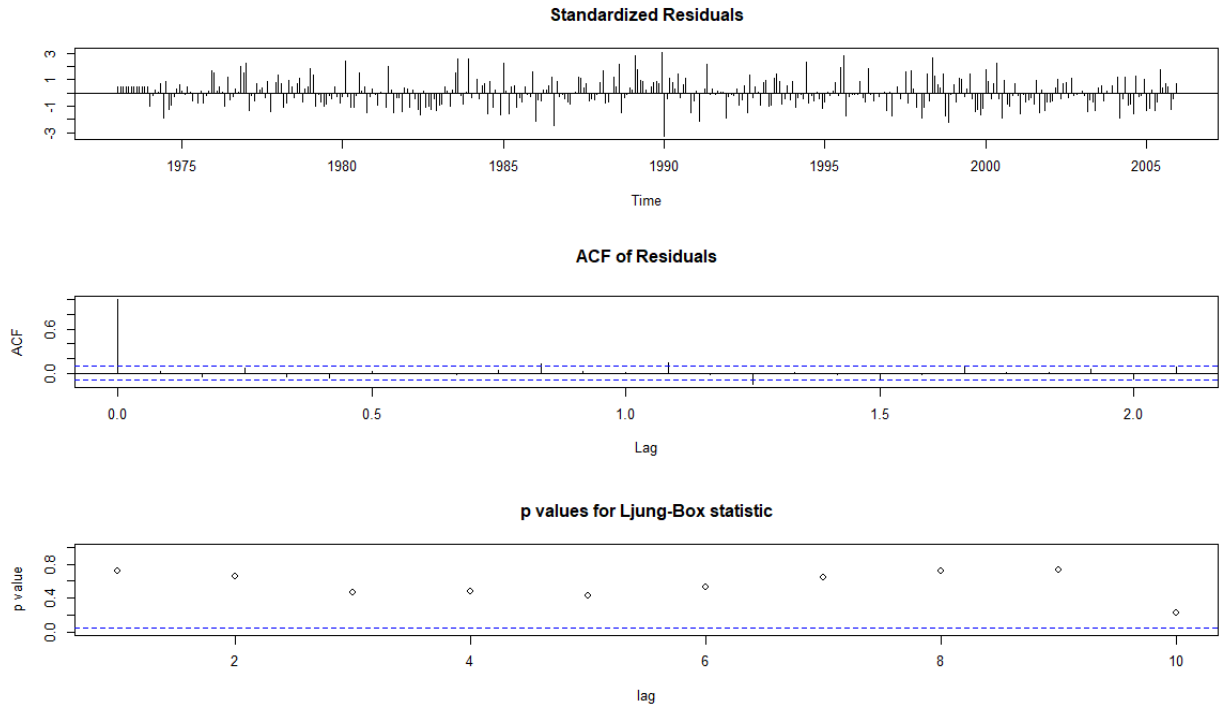

Figure 8: Residual Diagnostics

From the above plots we can see that:

1. The Standardized Residuals are randomly scattered around zero, thus homoscedastic.
2. There are no significant autocorrelations between the residuals at any lag.
3. The Ljung-Box statistic suggests that the residuals are Independent.

From the residual diagnostics we inferred that the forecast would be accurate and will not be significantly affected by violated assumptions.

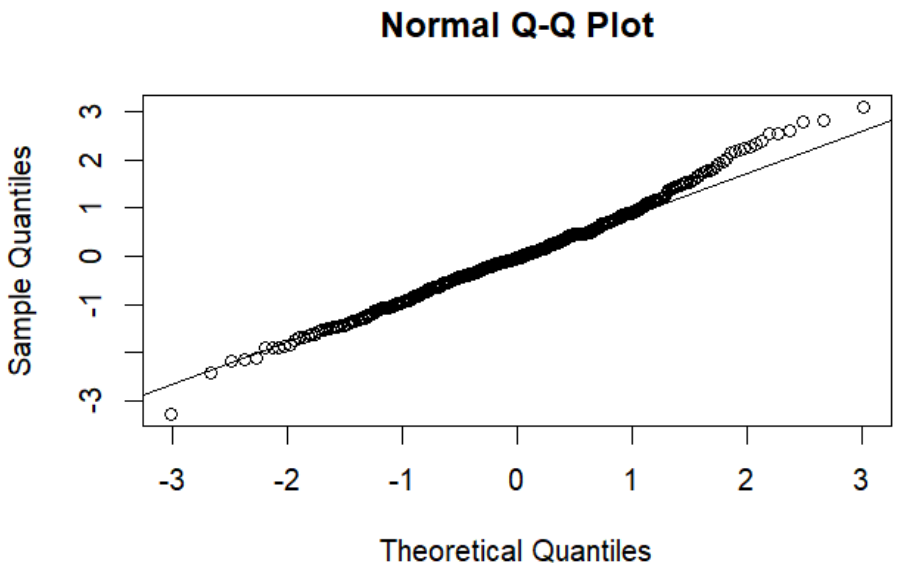

Figure 9: Quantile-Quantile Plot suggests normality

The QQ plot has some deviation toward the right, and the Shapiro-Wilk test outputs a value of 0.0338. However, the data was concluded to be Normal, due to the large size of the dataset.

We have overfitted the AR and MA part of the model to test for significant AR3 and MA2 coefficients as this would suggest the chosen model might be incorrect.

AR Order +1

```
ARIMA(3,0,1)(0,0,1)[12] with non-zero mean

Coefficients:

         ar1      ar2     ar3      ma1

      1.3878  -0.4505  0.0447  -0.8127

s.e.  0.0806   0.0896  0.0570   0.0589

         sma1    mean

      -0.8512  0.0248

s.e.   0.0312  0.0023
```

MA Order +1

```
ARIMA(2,0,2)(0,0,1)[12] with non-zero
mean

Coefficients:

         ar1      ar2      ma1      ma2

      1.2198  -0.2417  -0.6400  -0.1334

s.e.  0.2280   0.2177   0.2303   0.1473

         sma1    mean

      -0.8498  0.0248

s.e.   0.0316  0.0023
```

Figure 10: Testing for overfitting with 95% confidence

As the AR3 and MA2 components are insignificant we have concluded that our chosen model is correct for this data, and will yield reliable predictions.

Forecasting

The ARIMA(2,0,1)(0,0,1)[12] model was used to forecast monthly electricity production values five years ahead from the end of the data in December 2005. Figure 11 shows a plot of the data with the predicted values, as well as a 95% confidence interval. It can be concluded that the prediction exhibits the same trend as the data, best seen in Figure 12. There is no reason to believe that there would be any drastic actual changes in five years ahead. Thus, this prediction is reliable.

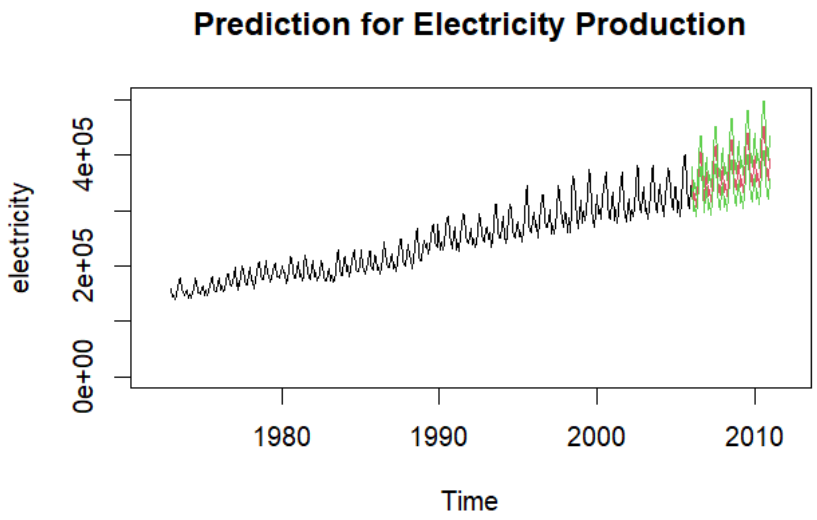

Figure 11: Prediction of the time series five years ahead.

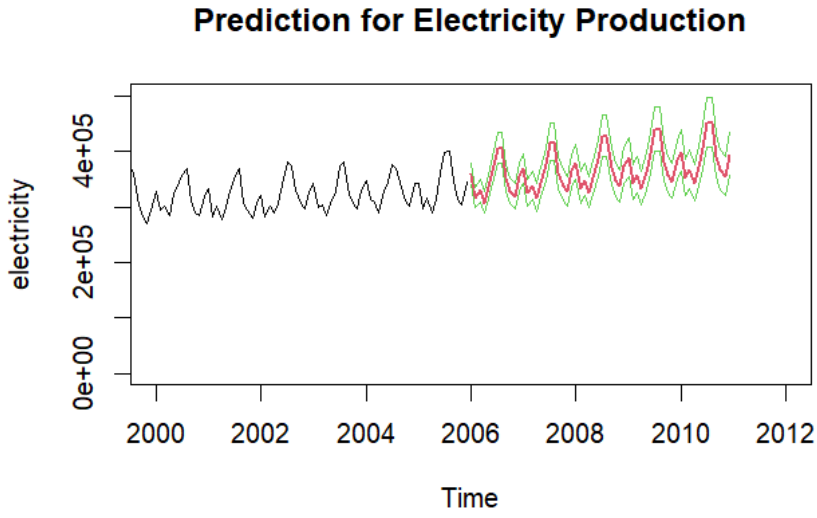

Figure 12: Prediction of the time series five years ahead, zoomed in on the predicted values to view seasonality.

Conclusion

By plotting the electricity dataset we found that there was a deterministic linear trend with some seasonality. So in order to perform the analysis we have log transformation of the data and performed ADF, PP and KPSS tests and found that the log transformed data was non-stationarity but we got rid of the variance in the data. So to address this non-stationarity we have performed seasonal differencing and performed ADF, PP and KPSS tests again and found that this data after log transformation and seasonal differencing is Stationary. By using ACF, PACF and EACF we were able to select a seasonal ARIMA model and for this model we performed model diagnostics to determine that there were no significant residuals and they were Independent, homoscedastic, and stationary.

To answer the research questions posted earlier. We were able to fit seasonal ARIMA model ARIMA(2,0,1)(0,0,1)[12] with seasonal differencing at lag 12 and with a period of 12. And we were able to forecast the electricity production with a 95% confidence interval for a period of 5 years starting from January 2006.

References:

1. Energy, J. (2022, November 21). The History of Electricity (& the Future of Electric Power). Just Energy. https://justenergy.com/blog/history-of-electricity-electric-power/
2. Lantero, A. (2014, November 18). The War of the Currents: AC vs. DC Power. Energy.gov. https://www.energy.gov/articles/war-currents-ac-vs-dc-power
3. U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. (n.d.). Www.eia.gov. Retrieved April 25, 2023, from https://www.eia.gov/totalenergy/data/browser/index.php?tbl=T01.01#/?f=M

R Code

```
library(TSA)
data(electricity)
elec <- electricity
plot(elec, main='Monthly US Electricity Production',
    ylab='Millions of Kwh',xlab='Year', type='l',
    ylim=c(0,400000))
plot(log(elec), main='Log Monthly US Electricity Production',
    ylab='Millions of Kwh',xlab='Year', type='l')
tseries::adf.test(elec)

# Log Transform #
elecl <- log(elec)
tseries::adf.test(elecl)
tseries::pp.test(elecl)
tseries::kpss.test(elecl)
acf(ts(elecl))
pacf(ts(elecl))

# Seasonal Difference #
elec.logd <- diff(log(elec),lag=12)
plot(elec.logd,main='Seasonal Difference')

acf(ts(elec.logd),lag.max=60,main='ACF of Seasonal Difference')
pacf(ts(elec.logd),lag.max=60,main='PACF of Seasonal Difference')
eacf(ts(elec.logd))

tseries::adf.test(elec.logd)
tseries::pp.test(elec.logd)
tseries::kpss.test(elec.logd)

# Selecting a Model
library(forecast)
auto.arima(elec.logd) #ARIMA(2,0,1)(2,0,1)[12]
Arima(elec.logd,order=c(2,0,1), ####
     seasonal=c(0,0,1))
Arima(elec.logd,order=c(1,0,3),
     seasonal=c(0,0,1))
Arima(elec.logd,order=c(2,0,2),
     seasonal=c(0,0,1))
```

```
Arima(elec.logd,order=c(1,0,2),
     seasonal=c(0,0,1))
Arima(elec.logd,order=c(1,0,1),
     seasonal=c(1,0,1))

model <- Arima(y=log(elec),order=c(2,0,1),
            seasonal=c(0,1,1),xreg=1:length(elec))

# Model Diagnostics
tsdiag(model)
qqnorm(rstandard(model))
qqline(rstandard(model))
shapiro.test(rstandard(model))

# Overfitting
Arima(elec.logd,order=c(3,0,1),
     seasonal=c(0,0,1))
Arima(elec.logd,order=c(2,0,2),
     seasonal=c(0,0,1))

# Prediction
pred <- predict(model,n.ahead=60,newxreg=length(elec)+1:60)
pr <- pred$pred
uci <- pr+2*pred$se
lci <- pr-2*pred$se
ymin <- min(c(as.vector(lci),elec))-100
ymax <- max(c(as.vector(uci),elec))+100000

plot(elec,ylim=c(ymin,ymax),xlim=c(1972,2012),
    main='Prediction for Electricity Production')
lines(exp(pr),col=2,lwd=1.5)
lines(exp(lci),col=3)
lines(exp(uci),col=3)

plot(elec,ylim=c(ymin,ymax),xlim=c(2000,2012),
    main='Prediction for Electricity Production')
lines(exp(pr),col=2,lwd=2)
lines(exp(lci),col=3)
lines(exp(uci),col=3)
```