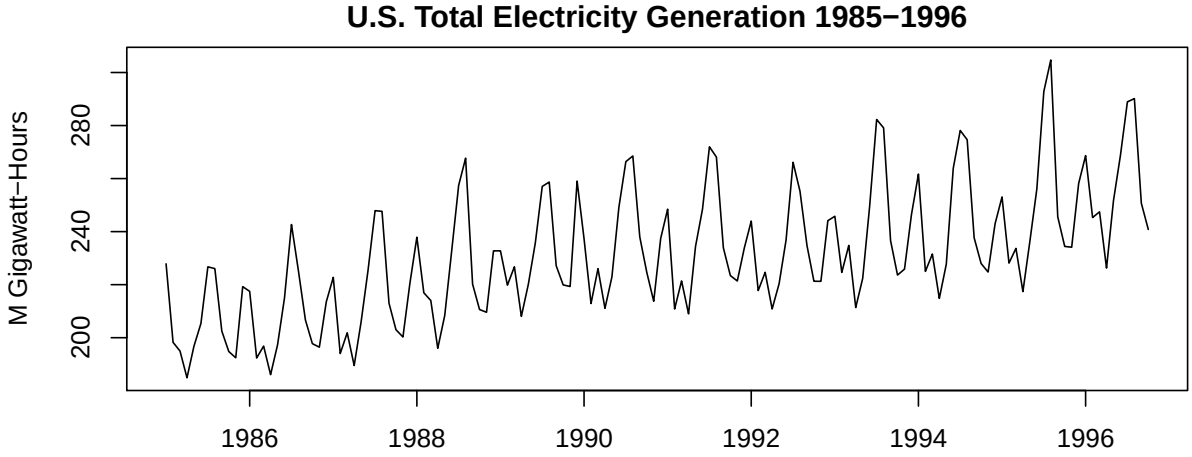
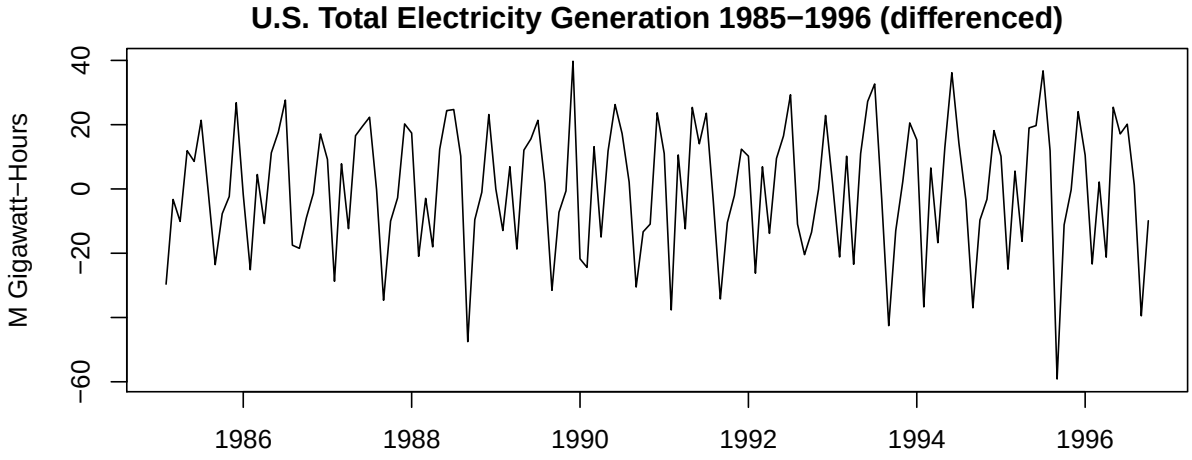
Forecasting Total U.S. Electricity Generation

David Noonan

This is an exercise in time series analysis, forecasting the total electricity generation in the United States. Since this is an exercise, I am using old data to forecast a time period that has already happened (but without looking). The data consists of monthly generation totals from January 1985 to October 1996 in millions of gigawatt-hours. We will explore the dataset, fit a model, and ultimately make a 12 month forecast of electricity generation for the year 1997. Below is a plot of the dataset.

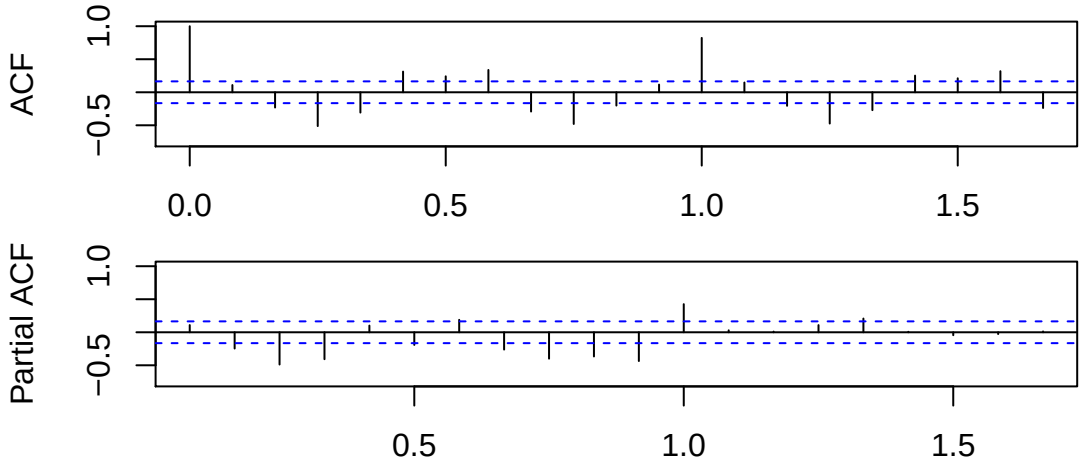


Notice that the series is not stationary. There is a consistent upward trend. Beyond the lack of stationarity, there is a seasonal component. Each year, we see a pattern of three major spikes. We can model this series with SARIMA, but first we need to transform the dataset to achieve stationarity. We can remove the trend by differencing. Below is a plot of the differenced series:

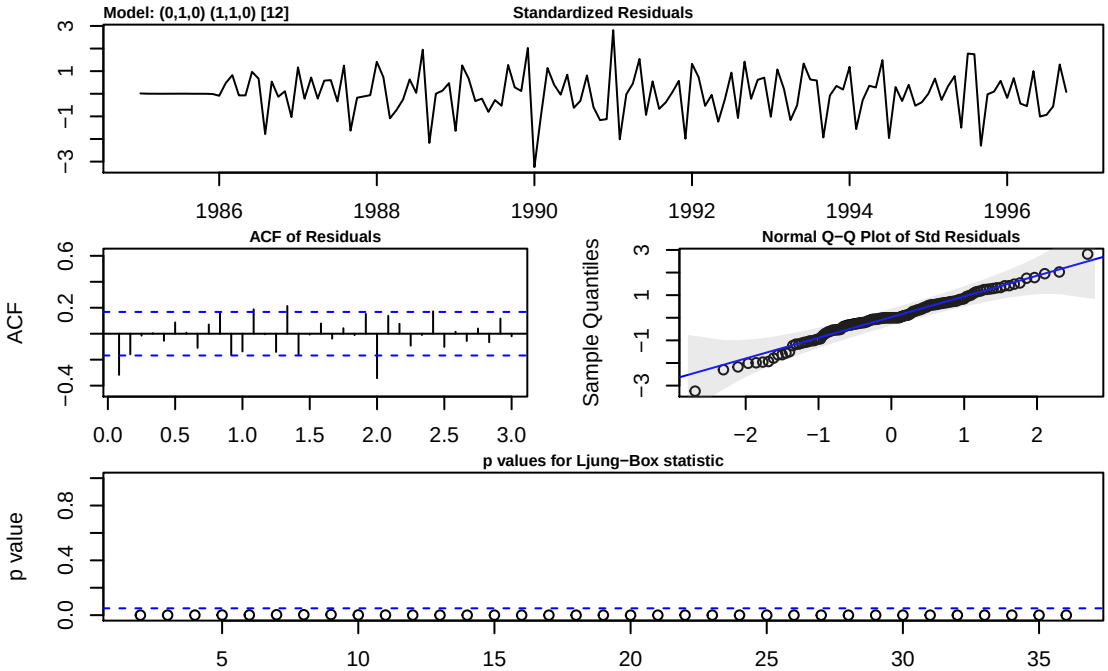


This series looks much closer to stationarity. We removed the trend, and achieved an apparent constant variance. Now we can begin building a model.

Our first step in model building is to look at the auto-correlation and partial-auto-correlation plots of our dataset to get an idea of the model parameters we should use:

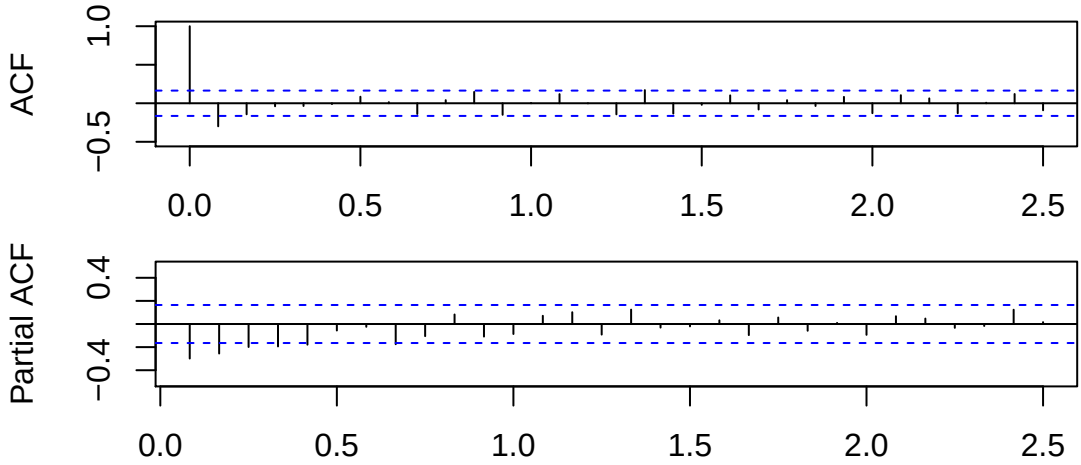


The ACF plot shows strong peaks at 1s, 2s, 3s... which is a slowly decaying seasonal effect. This may be an indication of seasonal nonstationarity, where the series is almost periodic by the seasons. We can remove this effect with seasonal differencing. What is more, the PACF plot cuts off after 1s (corresponding to 1 year). This suggests a SARI model of seasonal order P = 1 may be appropriate. Below is a set of diagnostic plots for the SARI model fit to this data, with seasonal differencing D = 1, and seasonal period of 12 months:



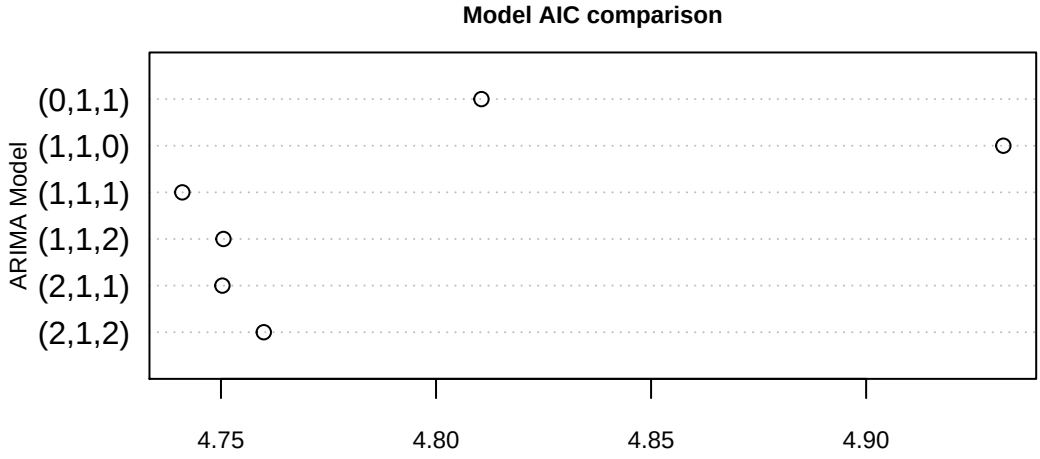
The diagnostic plots are encouraging. The residual time plot does not show any obvious patterns, and there are no outliers greater than 3 standard deviations from zero. The residual distribution in the Q-Q plot is very close to normal, but with slightly heavy tails. Two concerns are the residual ACF plot, which shows autocorrelation in the first lag, as well as the second seasonal lag. This may indicate a seasonal moving average component is necessary. Finally, the p-values of the Ljung-Box statistics indicate some departure from the independence assumption of the residuals.

Next, we look a the ACF and PACF plots of the residuals from a model with a seasonal moving average Q = 1 component added.



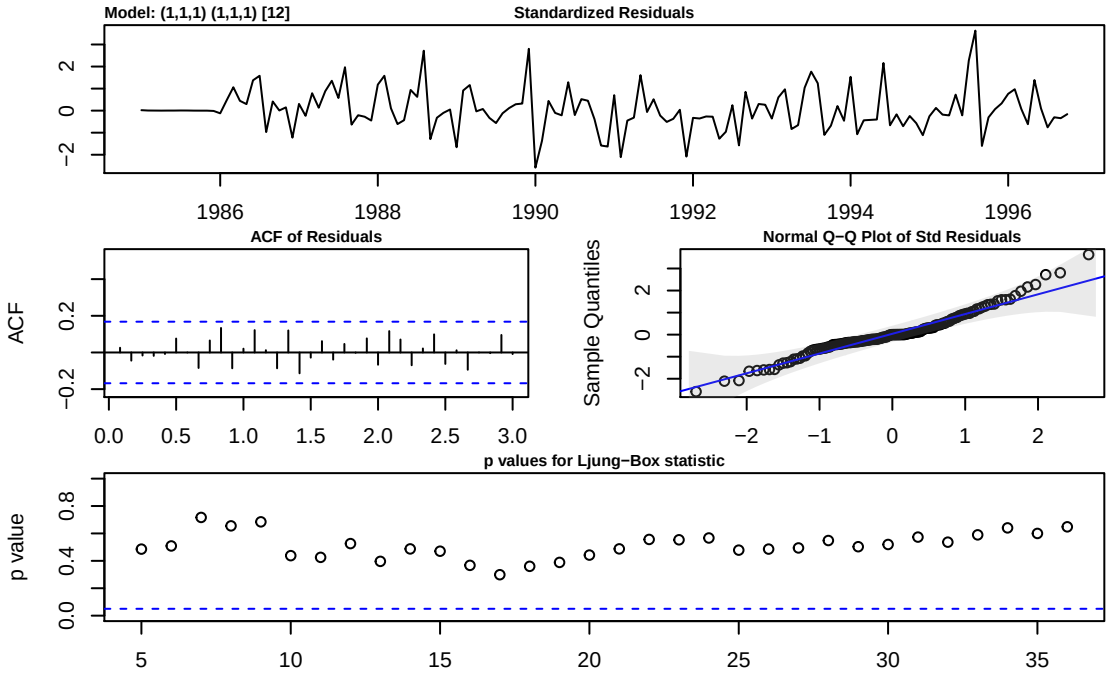
Here, both plots appear to be tailing off after one lag. This indicates a possible ARIMA process with orders $p > 0$ and $q > 0$. This suggests that we can add autoregressive and moving average parameters to the model. We will consider a few models, ordered by number of parameters, and compare them with information criterea: ARIMA$(2, 1, 2)$, ARIMA$(2, 1, 1)$, ARIMA$(1, 1, 2)$, ARIMA$(1, 1, 1)$, ARIMA$(1, 1, 0)$, and ARIMA$(0, 1, 1)$, each with seasonal component $(1, 1, 1)_{12}$.

Below is a dot chart comparison of the Akaike Information Critera (AIC) for each model fit:



In this plot we see that model 4, which is ARIMA$(1, 1, 1) \times (1, 1, 1)_{12}$ has the least AIC value. If the model diagnostics are good, then this could be a good model for forecasting the series. Next we will look at the diagnostic plots for this model fit.

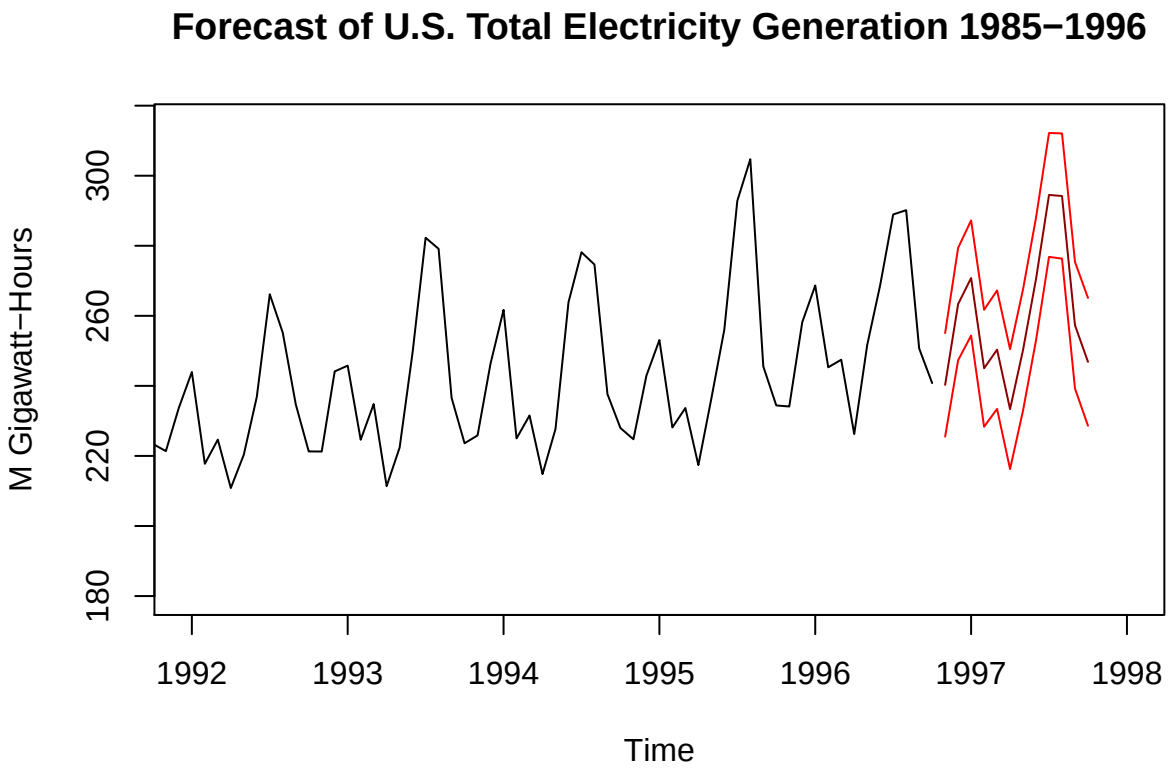
Below is the diagnostic plot for the model ARIMA$(1,1,1) \times (1,1,1)_{12}$:



The residual time plot shows no obvious pattern, and none of the residuals lie more then three standard devations from zero. The ACF plot of residuals shows no significant autocorrelation. The normal QQ plot looks decent, but with slightly heavy tails. Finally, the Ljung-Box statistic p-values indicate no departure from the independence assumption. This model looks good from the diagnostics, so we can move on to forecasting.

Forecasting

Now we will forecast the series 12 months into the future. For this we will use the **sarima.for** function from the astsa package in R. Below is a plot of the forecast, with a 99% confidence interval for prediction (red):



Overall, the forecast looks believable. The pattern of three spikes per year continues into the year we are forecasting, as does the overall upward trend. According to the model, U.S. Energy production will continue in its upward trend, with a similar seasonal pattern to previous years.

Code

```
library(dplyr)
library(astsa)
generation <- read.table(file = "generation.csv", header = TRUE) %>% #read data

# convert dataset to a timeseries object
ts(, start = c(1985,1), end = c(1996,10), frequency = 12)
par(mar = c(2,4,2,2))
plot.ts(generation, main = "U.S. Total Electricity Generation 1985-1996",
        ylab = "M Gigawatt-Hours")

# plot differenced timeseries
par(mar = c(2,4,2,2))
plot(diff(generation),main = "U.S. Total Electricity Generation 1985-1996 (differenced)",
     ylab = "M Gigawatt-Hours")

acfpar <- par(mfrow = c(2,1), mar = c(2,4,1,2), xlab = NA)
invisible(acfpar)
acf(diff(generation), lag.max = 20, main = NA, ylim = c(-.75, 1))  # acf plot
pacf(diff(generation), lag.max = 20, main = NA, ylim = c(-.75, 1)) # pacf plot

par(cex.main = .75, mar = c(2, 4, 1, 2))
fit <- sarima(generation, 0,1,0,P = 1, D = 1, S = 12,details = FALSE) #SARI model

#SARI model P = 1 Q = 1
fitsma <- sarima(generation, 0,1,0,P = 1, D = 1, Q = 1, S = 12,details = FALSE)

# ACF and PACF plot of residuals from initial fit
par(mfrow = c(2,1), mar = c(2,4,1,2), xlab = NA)
acf(fitsma$fit$residuals, lag.max = 30, main = NA, ylim = c(-.5, 1))
pacf(fitsma$fit$residuals, lag.max = 30, main = NA, ylim = c(-.5, .5))

#Candidate SARIMA models
#p = 2, q = 2
fit1 <- sarima(generation, p =2, d = 1, q = 2,  P = 1, D = 1, Q = 1,
               S = 12, details = FALSE)
#p = 2, q = 1
fit2 <- sarima(generation, p =2, d = 1, q = 1,  P = 1, D = 1, Q = 1,
               S = 12, details = FALSE)
#p = 1, q = 2
fit3 <- sarima(generation, p =1, d = 1, q = 2,  P = 1, D = 1, Q = 1, S = 12,
               details = FALSE)
#p = 1, q = 1
fit4 <- sarima(generation, p =1, d = 1, q = 1,  P = 1, D = 1, Q = 1, S = 12,
               details = FALSE)
#p = 1, q = 0
fit5 <- sarima(generation, p =1, d = 1, q = 0,  P = 1, D = 1, Q = 1, S = 12,
               details = FALSE)
#p = 0, q = 1
fit6 <- sarima(generation, p =0, d = 1, q = 1,  P = 1, D = 1, Q = 1, S = 12,
               details = FALSE)
```

```r
#Cleveland Dot Chart of AIC values
par(mar=c(2,4,2,2), cex.main = .75, cex.lab = .75, cex.axis = .75)
fitbics <- c(fit1$AIC, fit2$AIC, fit3$AIC, fit4$AIC, fit5$AIC, fit6$AIC)
fitnames <- c("(2,1,2)", "(2,1,1)", "(1,1,2)", "(1,1,1)", "(1,1,0)", "(0,1,1)")
dotchart(fitbics, labels = fitnames, main = "Model AIC comparison",
         ylab = "ARIMA Model", xlab = "AIC" )

par(cex.main = .75, mar = c(2, 4, 1, 2))
finfit <- sarima(generation, p =1, d = 1, q = 1,  P = 1, D = 1, Q = 1,
                 S = 12, details = FALSE)

#forecast
forc <- sarima.for(generation, n.ahead = 12, p =1, d = 1, q = 1,  P = 1,
                   D = 1, Q = 1, S = 12)
#forcast plot
plot.ts(generation, col = 1:2,
        main = "Forecast of U.S. Total Electricity Generation 1985-1996",
        ylab = "M Gigawatt-Hours", xlim = c(1992,1998), ylim = c(180, 315))
lines(forc$pred, col = "darkred") #mean prediction line
lines(forc$pred + forc$se*qnorm(.99), col = "red") #upper limit prediction
lines(forc$pred - forc$se*qnorm(.99), col = "red") #lower limit prediction
```

Thanks to Prof. Kerr from California State University East Bay!