Data 624: Project 1: Part B

Angrand, Burke, Deboch, Groysman, Karr

October 19, 2019

### Part B - Forecasting Power

**Data: ResidentialCustomerForecastLoad-624.xlsx**

Part B consists of a simple dataset of residential power usage for January 1998 until December 2013. Your assignment is to model these data and a monthly forecast for 2014. The data is given in a single file. The variable ‘KWH’ is power consumption in Kilowatt hours, the rest is straight forward. Add these to your existing files above - clearly labeled.

**a.** Read in File

temp = tempfile(fileext = ".xlsx")  
dataURL <- "https://raw.githubusercontent.com/mburke65/CUNY\_Data624/master/ProjectFolder/Provided\_Files/ResidentialCustomerForecastLoad-624.xlsx"  
download.file(dataURL, destfile=temp, mode='wb')  
  
power.data <- readxl::read\_excel(temp, sheet =1)

**b.** EDA Analysis

- Check/Fill in null values   
- Convert to time series  
- Graph the monthly data   
 - General plot & seasonal plot: seasonality can be observed in the below plot. There are spikes each year from May to August (air conditioning?) and again in December (holiday season?). There is a slight dip in Jul 2010 maybe due to an unseasonably cold month.  
 - Seasonal Box Plot: provides a similar visual to the seasonal plot with usage spikes in the summer months and December. IT also highlights the flucuations in consumption within each month.   
 - Decomposition components graph: this plot again shows that there is a general upwards trend in the data with an observed outlier in July 2010.

head(power.data)

## # A tibble: 6 x 3  
## CaseSequence `YYYY-MMM` KWH  
## <dbl> <chr> <dbl>  
## 1 733 1998-Jan 6862583  
## 2 734 1998-Feb 5838198  
## 3 735 1998-Mar 5420658  
## 4 736 1998-Apr 5010364  
## 5 737 1998-May 4665377  
## 6 738 1998-Jun 6467147

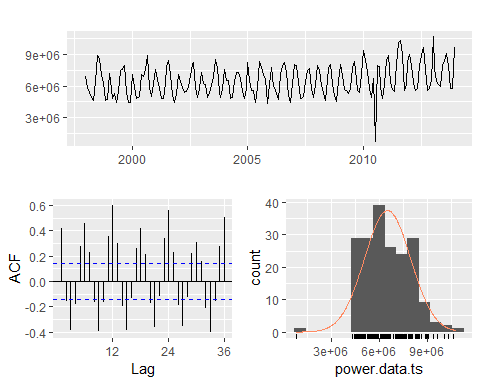
summary(power.data)

## CaseSequence YYYY-MMM KWH   
## Min. :733.0 Length:192 Min. : 770523   
## 1st Qu.:780.8 Class :character 1st Qu.: 5429912   
## Median :828.5 Mode :character Median : 6283324   
## Mean :828.5 Mean : 6502475   
## 3rd Qu.:876.2 3rd Qu.: 7620524   
## Max. :924.0 Max. :10655730   
## NA's :1

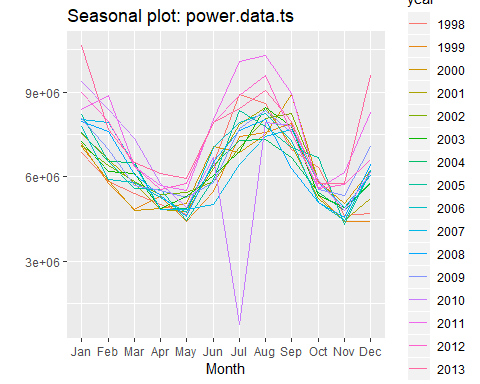
print(paste("Check for nulls: ",sum(is.na(power.data)), " Row of Nulls"))

## [1] "Check for nulls: 1 Row of Nulls"

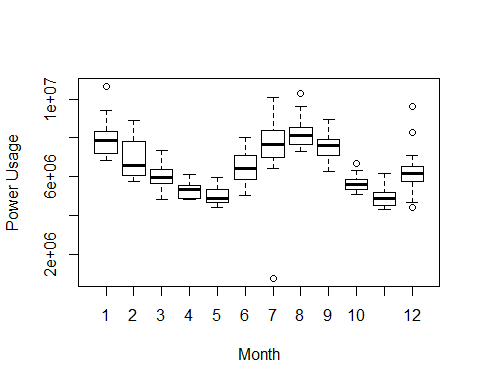
power.data[is.na(power.data)] <- median(power.data$KWH,na.rm = TRUE)  
power.data.ts <-ts(power.data[,"KWH"],start = c(1998,1),frequency = 12)  
ggtsdisplay(power.data.ts, points = FALSE, plot.type = "histogram")



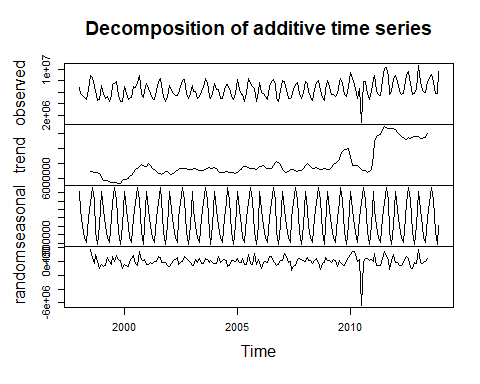
ggseasonplot(power.data.ts)



boxplot(power.data.ts~cycle(power.data.ts),xlab="Month", ylab = "Power Usage")



plot(decompose(power.data.ts))



**C.** Model 1: Arima W/ Box-Cox Transformation

- Replace outlier with tsoutlier suggestion (utilizes a box-cox transformation)  
- Use an auto arima model on the box-cox adjusted data  
 - Suggested model: ARIMA(0,0,3)(2,1,0)[12] with drift. RSME(595389) & AICc (5332.67)  
- Check the residuals to make sure the model is satisfactory:   
 - ACF /PACF Plots: the residual appears normal residuals mostly around 0, suggesting stationarity of the residuals   
 - The Box Ljung tests presents a p-value of 0.6951 which indicates white noise  
- Forecast 2014 power values & plot forecasted values

#outlier detection/suggestion/replacement  
find.outlier<- tsoutliers(power.data.ts, iterate = 2, lambda = "auto")  
power.data.ts.bc<- power.data.ts  
power.data.ts.bc[find.outlier$`index`[1]] <- find.outlier$replacements[1]  
print(paste("Suggested/Implemented Change for Outlier: ",power.data.ts.bc[151], " Original Value",power.data.ts[151]))

## [1] "Suggested/Implemented Change for Outlier: 7757388.48810024 Original Value 770523"

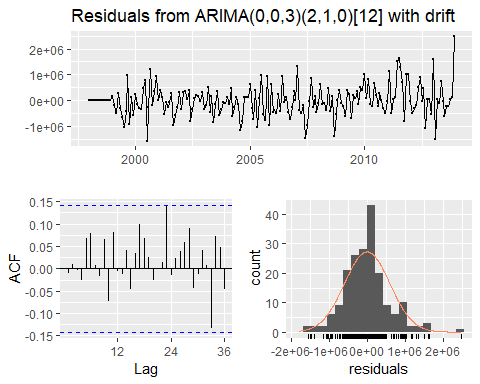
#auto arima model  
power.model <- auto.arima(power.data.ts.bc, seasonal = TRUE, stepwise = FALSE)  
summary.arima<- summary(power.model)

## Series: power.data.ts.bc   
## ARIMA(0,0,3)(2,1,0)[12] with drift   
##   
## Coefficients:  
## ma1 ma2 ma3 sar1 sar2 drift  
## 0.3492 0.0587 0.2303 -0.7222 -0.4251 9027.233  
## s.e. 0.0788 0.0892 0.0741 0.0765 0.0784 3057.838  
##   
## sigma^2 estimated as 3.912e+11: log likelihood=-2659.01  
## AIC=5332.02 AICc=5332.67 BIC=5354.37  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -8181.906 595389 434520.9 -0.8060555 6.610412 0.6977491  
## ACF1  
## Training set -0.01026567

summary.arima

## ME RMSE MAE MPE MAPE MASE  
## Training set -8181.906 595389 434520.9 -0.8060555 6.610412 0.6977491  
## ACF1  
## Training set -0.01026567

#check residuals  
checkresiduals(power.model)

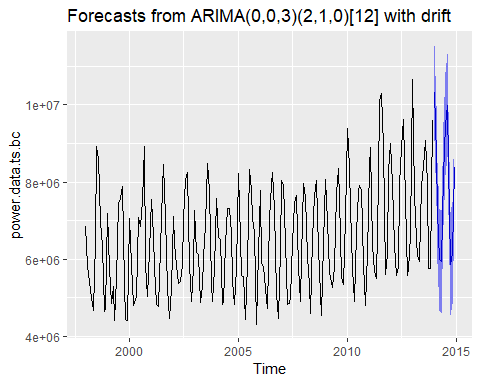


##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(0,0,3)(2,1,0)[12] with drift  
## Q\* = 14.513, df = 18, p-value = 0.6951  
##   
## Model df: 6. Total lags used: 24

#forecast model @ 95%  
forecast.power <- forecast(power.model, level = c(95), h =12)  
forecast.power

## Point Forecast Lo 95 Hi 95  
## Jan 2014 10312755 9086940 11538570  
## Feb 2014 8685531 7387130 9983933  
## Mar 2014 7203085 5902687 8503482  
## Apr 2014 6000251 4669575 7330927  
## May 2014 5941905 4611229 7272581  
## Jun 2014 8204931 6874255 9535607  
## Jul 2014 9501418 8170742 10832094  
## Aug 2014 9992966 8662290 11323642  
## Sep 2014 8493959 7163283 9824635  
## Oct 2014 5871672 4540996 7202348  
## Nov 2014 6154352 4823676 7485028  
## Dec 2014 8381806 7051130 9712482

autoplot(forecast.power)



**D.** Model 2: ETS W/ Box-Cox Transformation

- The ets function automatically selects the best method for forecasting data. the ets function selected ETS(M,N,M) exponential smoothing:  
 - The first letter denotes the error type: multiplicative errors  
 - The second letter denotes the trend type: no trend  
 - The third letter denotes the season type: multiplicative seasonality  
- Utilize the transformed data & ETS model  
- Model Results: RMSE (630869.7) & AICc (6148.032)  
- Check the residuals to make sure the model is satisfactory:   
 - ACF /PACF Plots: the residual appears normal residuals mostly around 0, suggesting stationarity of the residuals   
 - The Box Ljung tests presents a p-value of 0.0002921 which may indicate that there's dependency issues with the lags  
 -

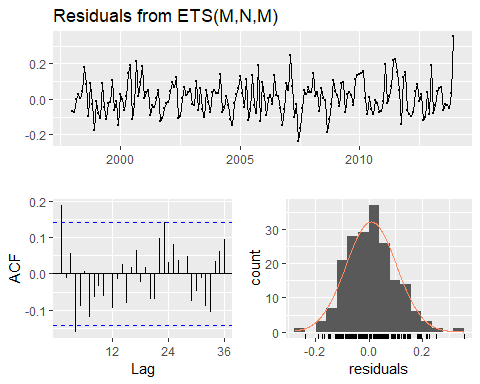
#model w/ previously transformed data  
power.model.ets <- ets(power.data.ts.bc)  
summary.ets<- summary(power.model.ets)

## ETS(M,N,M)   
##   
## Call:  
## ets(y = power.data.ts.bc)   
##   
## Smoothing parameters:  
## alpha = 0.1206   
## gamma = 0.203   
##   
## Initial states:  
## l = 6188160.6435   
## s = 0.9017 0.755 0.9295 1.223 1.2676 1.2298  
## 1.0165 0.7614 0.8029 0.8903 1.029 1.1935  
##   
## sigma: 0.0971  
##   
## AIC AICc BIC   
## 6145.305 6148.032 6194.167   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 51458.11 630869.7 482886.1 0.04287257 7.292037 0.7754135  
## ACF1  
## Training set 0.2096574

summary.ets

## ME RMSE MAE MPE MAPE MASE  
## Training set 51458.11 630869.7 482886.1 0.04287257 7.292037 0.7754135  
## ACF1  
## Training set 0.2096574

#check residuals  
checkresiduals(power.model.ets)

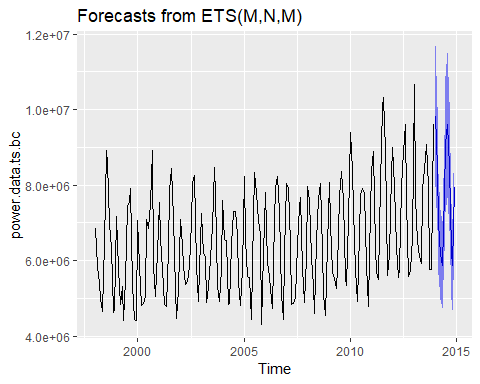


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,M)  
## Q\* = 32.819, df = 10, p-value = 0.0002921  
##   
## Model df: 14. Total lags used: 24

#forecast model @ 95%  
forecast.power.ets <- forecast(power.model.ets, level = c(95), h =12)  
forecast.power.ets

## Point Forecast Lo 95 Hi 95  
## Jan 2014 9825114 7955694 11694534  
## Feb 2014 8460361 6838842 10081879  
## Mar 2014 6974291 5627960 8320623  
## Apr 2014 6167737 4968643 7366830  
## May 2014 5886368 4733958 7038779  
## Jun 2014 7783200 6248904 9317496  
## Jul 2014 9070884 7270556 10871212  
## Aug 2014 9599368 7681334 11517403  
## Sep 2014 8501578 6791613 10211542  
## Oct 2014 6241977 4978271 7505684  
## Nov 2014 5885873 4686553 7085194  
## Dec 2014 7933193 6306380 9560005

autoplot(forecast.power.ets)



**E.** Model 3: STLF

- STLF model will be the third model as it provides the user more control and can be robust when dealing with outliers. the STLF utilizes a local weighted regression to fit the points (Loess smoothing) and forecast future values.  
- Model summary: RMSE (843670.1) & AICc (6255.445)  
- Check residuals:   
 - ACF/PACF: most lags are within the error bounds, suggesting stationarity of the residuals   
 - Box Ljung:p-value = 0.1457 which indicates white noise  
- Forecast 2014 power values & plot forecasted values

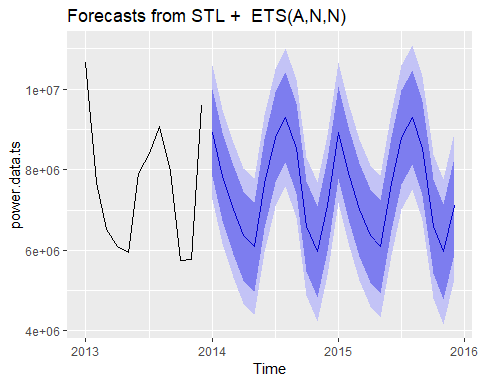
power.model.stl <- stlf(power.data.ts, s.window='periodic', robust=TRUE)   
summary.stl<- summary(power.model.stl)

##   
## Forecast method: STL + ETS(A,N,N)  
##   
## Model Information:  
## ETS(A,N,N)   
##   
## Call:  
## ets(y = x, model = etsmodel, allow.multiplicative.trend = allow.multiplicative.trend)   
##   
## Smoothing parameters:  
## alpha = 0.0892   
##   
## Initial states:  
## l = 6317161.2015   
##   
## sigma: 848098.8  
##   
## AIC AICc BIC   
## 6255.318 6255.445 6265.090   
##   
## Error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 69834.05 843670.1 512067.7 -4.243142 12.03155 0.7316422  
## ACF1  
## Training set 0.209786  
##   
## Forecasts:  
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## Jan 2014 8919230 7832347 10006112 7256987 10581473  
## Feb 2014 7833393 6742199 8924586 6164556 9502230  
## Mar 2014 7005154 5909665 8100642 5329749 8680558  
## Apr 2014 6358706 5258940 7458473 4676759 8040654  
## May 2014 6086344 4982317 7190371 4397880 7774808  
## Jun 2014 7653295 6545023 8761567 5958339 9348251  
## Jul 2014 8801193 7688692 9913693 7099770 10502616  
## Aug 2014 9301580 8184867 10418293 7593714 11009445  
## Sep 2014 8524478 7403568 9645387 6810194 10238761  
## Oct 2014 6593256 5468165 7718347 4872577 8313934  
## Nov 2014 5961092 4831835 7090349 4234043 7688141  
## Dec 2014 7113767 5980360 8247174 5380371 8847164  
## Jan 2015 8919230 7781688 10056772 7179509 10658950  
## Feb 2015 7833393 6691730 8975055 6087371 9579415  
## Mar 2015 7005154 5859386 8150921 5252853 8757454  
## Apr 2015 6358706 5208848 7508565 4600150 8117263  
## May 2015 6086344 4932409 7240279 4321553 7851135  
## Jun 2015 7653295 6495299 8811292 5882292 9424298  
## Jul 2015 8801193 7639149 9963237 7023999 10578386  
## Aug 2015 9301580 8135502 10467658 7518218 11084942  
## Sep 2015 8524478 7354380 9694575 6734968 10313987  
## Oct 2015 6593256 5419152 7767359 4797619 8388892  
## Nov 2015 5961092 4782996 7139188 4159350 7762834  
## Dec 2015 7113767 5931693 8295842 5305940 8921594

summary.stl

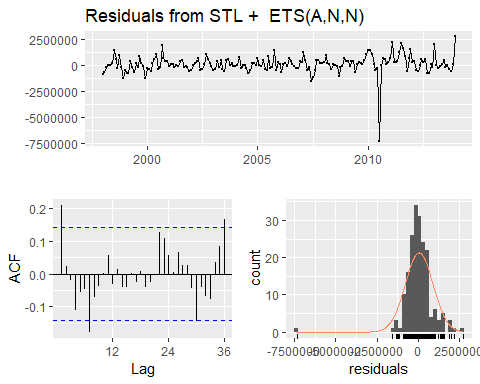
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## Jan 2014 8919230 7832347 10006112 7256987 10581473  
## Feb 2014 7833393 6742199 8924586 6164556 9502230  
## Mar 2014 7005154 5909665 8100642 5329749 8680558  
## Apr 2014 6358706 5258940 7458473 4676759 8040654  
## May 2014 6086344 4982317 7190371 4397880 7774808  
## Jun 2014 7653295 6545023 8761567 5958339 9348251  
## Jul 2014 8801193 7688692 9913693 7099770 10502616  
## Aug 2014 9301580 8184867 10418293 7593714 11009445  
## Sep 2014 8524478 7403568 9645387 6810194 10238761  
## Oct 2014 6593256 5468165 7718347 4872577 8313934  
## Nov 2014 5961092 4831835 7090349 4234043 7688141  
## Dec 2014 7113767 5980360 8247174 5380371 8847164  
## Jan 2015 8919230 7781688 10056772 7179509 10658950  
## Feb 2015 7833393 6691730 8975055 6087371 9579415  
## Mar 2015 7005154 5859386 8150921 5252853 8757454  
## Apr 2015 6358706 5208848 7508565 4600150 8117263  
## May 2015 6086344 4932409 7240279 4321553 7851135  
## Jun 2015 7653295 6495299 8811292 5882292 9424298  
## Jul 2015 8801193 7639149 9963237 7023999 10578386  
## Aug 2015 9301580 8135502 10467658 7518218 11084942  
## Sep 2015 8524478 7354380 9694575 6734968 10313987  
## Oct 2015 6593256 5419152 7767359 4797619 8388892  
## Nov 2015 5961092 4782996 7139188 4159350 7762834  
## Dec 2015 7113767 5931693 8295842 5305940 8921594

power.model.stl<- forecast(power.model.stl)  
autoplot(power.model.stl, 12)



checkresiduals(power.model.stl)

## Warning in checkresiduals(power.model.stl): The fitted degrees of freedom  
## is based on the model used for the seasonally adjusted data.



##   
## Ljung-Box test  
##   
## data: Residuals from STL + ETS(A,N,N)  
## Q\* = 28.969, df = 22, p-value = 0.1457  
##   
## Model df: 2. Total lags used: 24

**F.** Compare Model Results/Export Data

- After comparing the RMSE of in the accuracy test, the ARIMA model will be used as the final model due to the lower RSME and better prediction capabilities. The ARIMA model also has the lowest AICc score and best score from the Box Ljung tests.

rmse.list <- data.frame(list(accuracy(power.model)[2], accuracy(power.model.ets)[2], accuracy(power.model.stl)[2]))  
names(rmse.list)<- list('Arima', 'ETS','STL')  
rmse.list

## Arima ETS STL  
## 1 595389 630869.7 843670.1

**G.** Send Results to excel

- Send to a .csv file, will manually merge into the project's consolidated file for project submission

write.csv(forecast.power,"Power\_Forecasts\_ARIMA.csv")