Stat 638 Final Project

Hourly Energy Consumption

Guangting Zhou 2019.12.10

Summary

Energy consumption has been a very important part of our lives, we consume energy nearly everywhere and every time. Does the energy consumption has certain tendency? Is it predictable? All questions seem to be an interesting direction to approach for. To find the tendency and do some prediction on hourly energy consumption, I built some models including regression and time series seasonal ARIMA to do analyze. Based on the model validation, it turned out to show that specifically in this problem, the time series method seems to have significant advantages. The seasonal ARIMA model considering about the autocorrelation from the data itself and can possibly carry its memory and impacts to move on, which makes it a quite efficient model overall for the relatively stable data. My final model turn out to be the seasonal ARIMA model $SARIMA(1,1,0)*(0,1,2)_{12}$ for the hourly energy consumption by month in the United States. Here follows the detailed process for my project:

Introduction

Nowadays, our daily lives are full of electronic products. We consume electricity every day and moment, probably from waking up in the morning till falling asleep at night, the electricity energy consumption has been a very important part of our lives. Therefore, it rises up a question that how much energy can we consume hourly, does it has certain tendency or it is predictable. Sounds like a quite interesting question. Considering its possibly connections between time series, I finally focus my ideas on this topic for the time series final project.

Data Preprocessing

Finding the data sources:

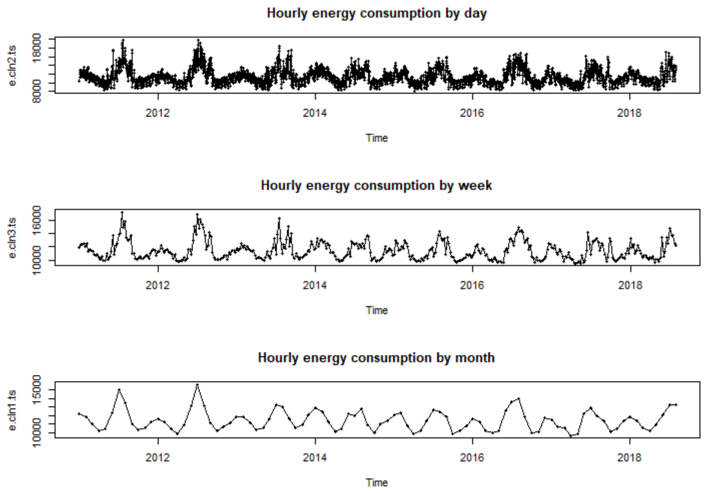
Since I've focused my topic on energy consumption, so I started to look through the websites to find datasets that can possibly be used. There're many online resource related to energy consumption, considering about the tidy and accuracy level, I finally chose the dataset from the Kaggle dataset website, focusing my analyzing dataset on the US hourly energy consumption from 2011 to 2018 recorded by one of the energy supplier ComEd. Here's the link: https://www.kaggle.com/robikscube/hourly-energy-consumption#COMED_hourly.csv. Considering that this dataset contains only the energy consumption value and time, the variables are too limited. Considering that the factors like weather or holidays may impact the energy consumption. I decided to add these variables into the dataset. So I looked through the website to find the historical records of climate data in the US: https://www.usclimatedata.com/. Abstracted the temperature data from 2011 to 2018, and applied these values into a new variable temperature. And by searching the yearly holidays by month, I added this as another new variable holiday in the dataset.

Data cleaning:

My original data was the hourly energy consumption in the US recorded by hour from Jan-01 01:00:00, 2011 to Aug-03 00:00:00, 2018 provided by ComEd, totally 66497 observations. Considering that the energy consumption level is not so time sensitive as stock or oil price, I decided to merge it into daily, weekly and monthly average records respectively. Also, by putting new variables into the transformed datasets, the datasets can therefore contain other factors, which would make it ready for regression analysis.

Descriptive and Exploratory Analysis

The datasets has been preprocessed, now it's ready to do further analysis. Here're the plots for hourly energy consumption by day, week and month as well as summary information for each variable:



```
> summary(e.cln1.ts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9626   10478   11298   11437   12132   15683 
> summary(energy.cln1$temp)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  42.80   54.75   68.75   68.23   82.42   93.40 
> summary(energy.cln1$holidays)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  1.0000  0.8152  1.0000  2.0000 
```

Based on the plot comparisons between different frequency, it looks that the frequency by month can be a good focusing point to show clear tendency. My following analysis will focus on the hourly energy consumption by month.

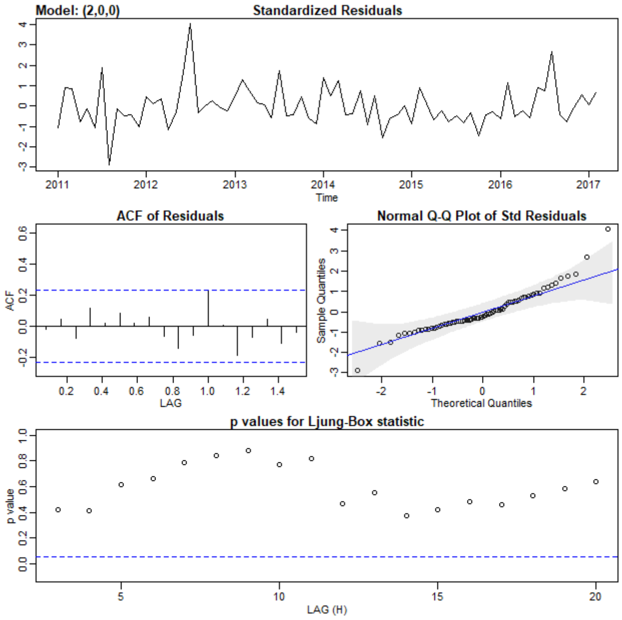
Model Building and Analysis

Regression with Autocorrelated Errors

Firstly, I considered about the regression method with autocorrelated errors. By putting the possibly impact factors into the regression model I got the following results:

```
> summary(fit <- lm(e.t.ts~trend + temp + temp2+ hd, na.action=NULL))

Call:
lm(formula = e.t.ts ~ trend + temp + temp2 + hd, na.action = NULL)

Residuals:
     Min       1Q   Median       3Q      Max
-1402.32  -340.93   -61.77   262.27  2471.27

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 37304.7292 74901.4723   0.498    0.620
trend         -13.4421    37.1886  -0.361    0.719
temp           29.7336     5.1339   5.792 1.89e-07 ***
temp2           5.7847     0.3824  15.128  < 2e-16 ***
hd            -19.2519   110.7330  -0.174    0.862
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 563.5 on 69 degrees of freedom
Multiple R-squared:  0.8022,    Adjusted R-squared:  0.7907
F-statistic: 69.94 on 4 and 69 DF,  p-value: < 2.2e-16
```

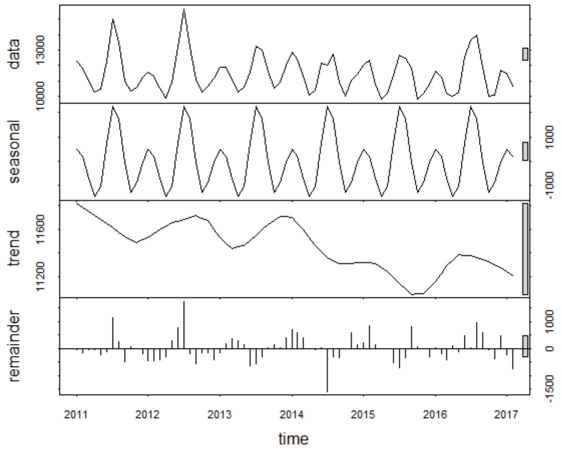
Based on the R output, it seems that the regression had large residuals. Then try to consider the residual as autocorrelated errors:



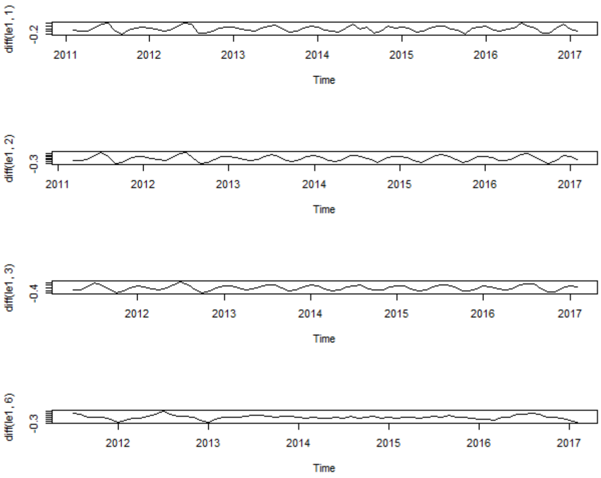
Therefore, after considering the autocorrelated error as regression error, we got the final model, with AIC=15.59265:

```
$ttable
            Estimate         SE t.value p.value
ar1           0.3358     0.1452  2.3124  0.0239
ar2          -0.2782     0.1447 -1.9220  0.0589
ar3          -0.0800     0.1455 -0.5497  0.5844
intercept 64865.0951 71239.0310  0.9105  0.3659
trend       -27.0833    35.3500 -0.7661  0.4463
temp         32.7302     5.2459  6.2392  0.0000
temp2         4.8266     0.8116  5.9471  0.0000
hd          117.7114    93.7211  1.2560  0.2136

$AIC
[1] 15.59265
```

Time Series SARIMA Models

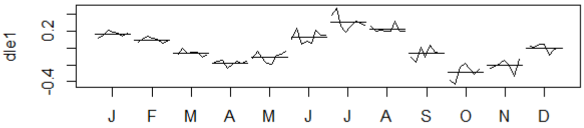


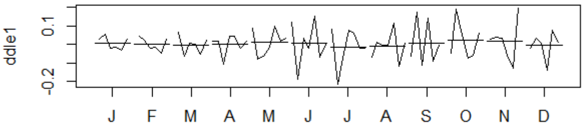
Based on the decomposition plot, it looks that the energy consumption data can possibly have a seasonal tendency. Thus I tried to build a SARIMA model.

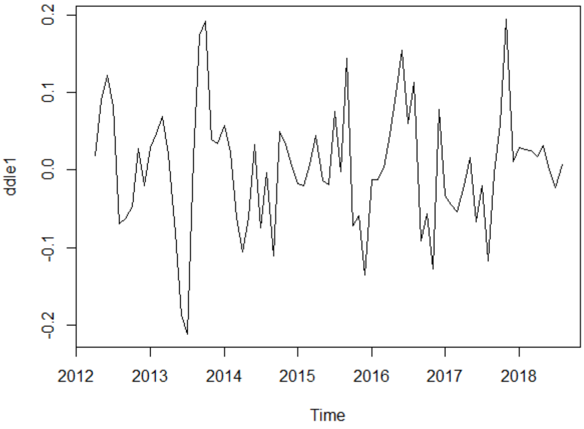


Model Diagnostics

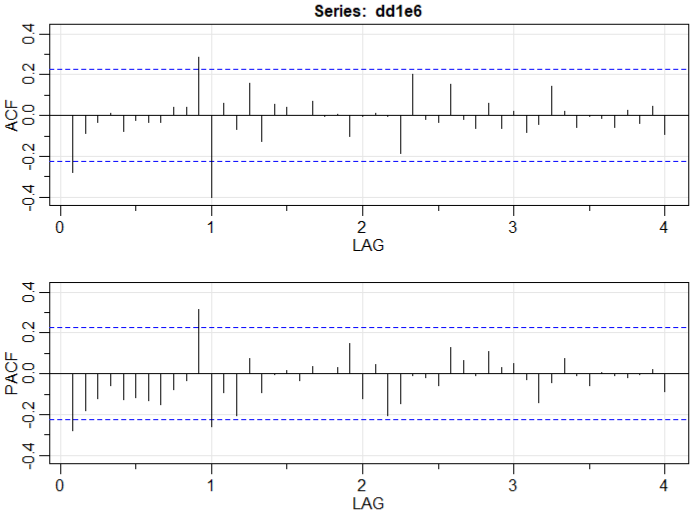
After the first and second difference, it seems that the first difference with 1 order, the second difference with 12 orders can be a proper transformation.



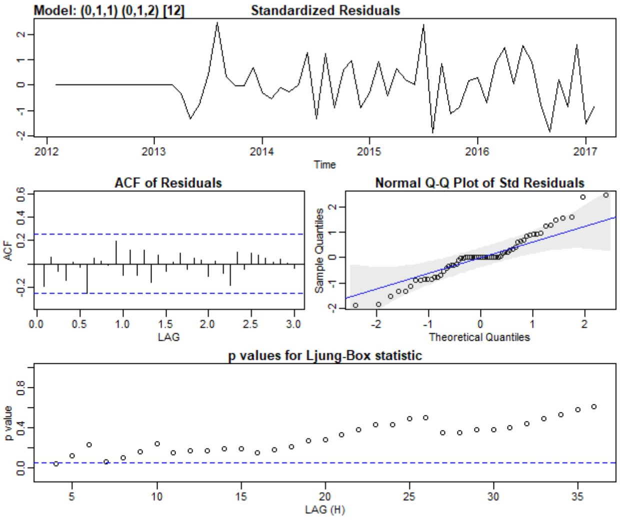


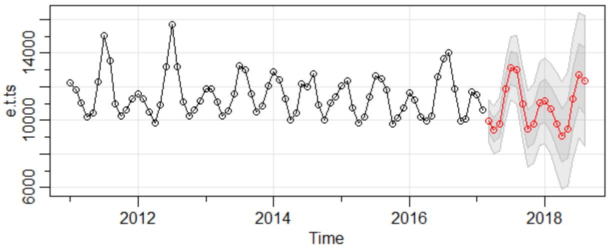


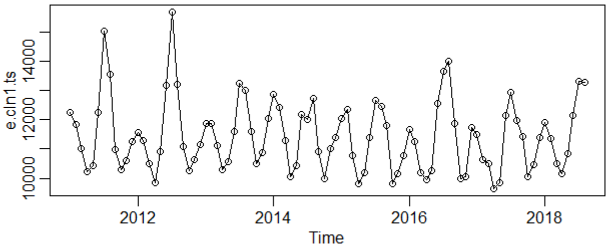
After the transformation, I tried to find information from the ACF and PACF plot, then built models:



Based on the acf and pacf, for seasonal, it can be P=0, Q=1 or 2, for non-seasonal, it can be p=0 or 1, q=0, 1, P=0 or 1, Q=0, 1 or 2. After considering about the AICs and residual conditions, it can be $(0, 1, 1) * (0, 1, 2)_{12}$ with AIC=-1.643136





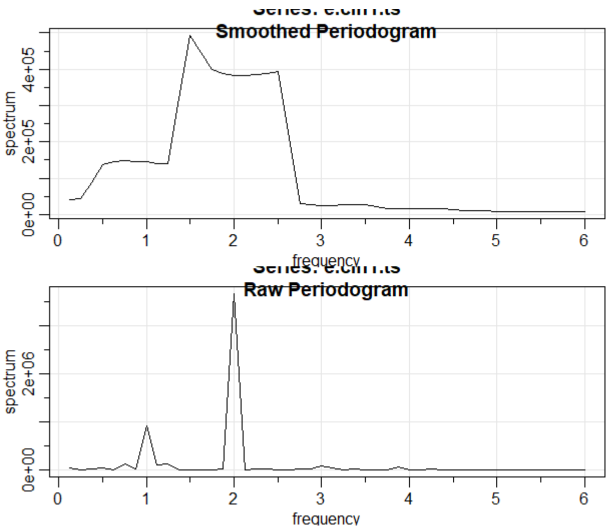


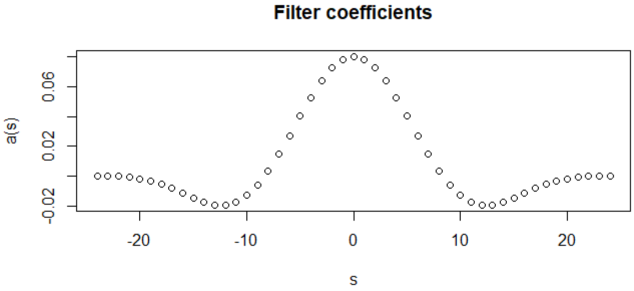
```
> sarima.for(e.t.ts, 18, 0,1,1, 0,1,2, 12 )
$pred
           Jan       Feb       Mar       Apr       May
      Jun       Jul       Aug       Sep
2017                          9982.722  9442.441  9798.087
11843.364 13098.198 12998.953 10953.749
2018 11159.383 10667.778  9802.818  9095.662  9480.349
11282.631 12684.495 12341.657
           Oct       Nov       Dec
2017  9505.348  9814.209 11007.699
2018

$se
           Jan       Feb       Mar       Apr       May
      Jun       Jul       Aug       Sep
2017                          627.5153  722.0222  805.5164
 881.1341  950.7565 1015.6173 1076.5774
2018 1292.2071 1340.2619 1509.3926 1605.8438 1696.8212
1783.1630 1865.5129 1944.3782
           Oct       Nov       Dec
2017 1134.2661 1189.1594 1241.6282
2018

```

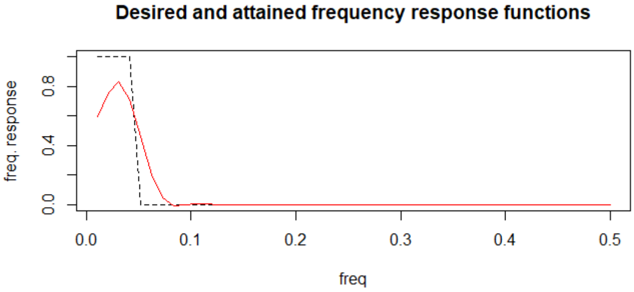
Model Validation and Performance Evaluation

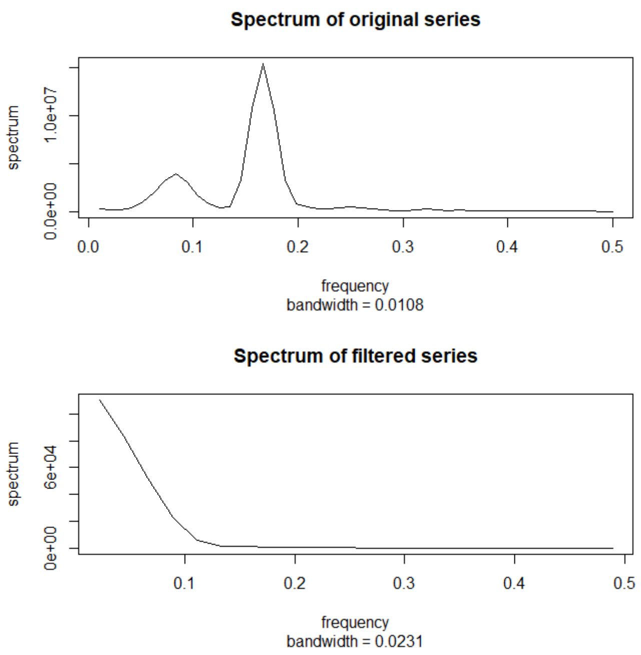
```
> accuracy(m1$pred, testing$consumption[-1])
                ME     RMSE      MAE      MPE     MAPE
Test set 491.5561 707.9002 625.5982 4.421993 5.529646
```

Spectral Analysis









```
> SigExtract(e.cln1.ts)
The filter coefficients are
          s          a(s)
 [1,]   0  8.000000e-02
 [2,]  -1  7.810597e-02
 [3,]  -2  7.258885e-02
 [4,]  -3  6.392528e-02
 [5,]  -4  5.285104e-02
 [6,]  -5  4.028083e-02
 [7,]  -6  2.721038e-02
 [8,]  -7  1.461279e-02
 [9,]  -8  3.341906e-03
[10,]  -9 -5.946299e-03
[11,] -10 -1.284528e-02
[12,] -11 -1.721763e-02
[13,] -12 -1.918221e-02
[14,] -13 -1.907220e-02
[15,] -14 -1.737134e-02
[16,] -15 -1.463808e-02
[17,] -16 -1.142862e-02
[18,] -17 -8.229104e-03
[19,] -18 -5.405759e-03
[20,] -19 -3.178390e-03
[21,] -20 -1.619119e-03
[22,] -21 -6.743622e-04
[23,] -22 -2.046001e-04
[24,] -23 -3.403312e-05
[25,] -24 -9.669615e-07
for s >=0; and a(-s) = a(s).
```

Interpretation

The tendency for hourly energy consumption data can overall follow a seasonal arima model $(0,1,1)*(0,1,2)_{12}$. While based on the spectral analysis, we can see that the series of hourly energy consumption are made up with a lot of $sine$ and $cosine$ waves. The actual points can be so complicate, while the overall trend can be represented by time series models like sarima.

Conclusion

Based on the built model, the hourly energy consumption can affected by so many factors, while overall, it can be putted into a SARIMA model. Where comparing with the regression model with autocorrelated errors, the SARIMA model can have significant advantage specifically for this dataset in fitting and forecasting, no matter in the efficiency, accuracy or the AIC. Though the actual values can be impacted by so many factors and be so complicative, the time series data itself may have its own memory, and such kind of memories like this hourly energy consumption are so powerful and stable, that there're so many other small impacts can only contribute limited level of influence waves to the whole wave. Thanks to all kinds of models and analysis methods, which make the explanation and prediction a lot easier. Now we can basically apply them into use, so that for a certain level of error, it can show us with a clear overall tendency, which so benefit us a lot in our work or study.

Appendix

```r
library(astsa)
energy <- read.csv(file.choose())
date <- as.Date(energy$Datetime)
energy$date <- date
library(tidyr)
energy.upd <- energy %>% separate(date, sep="-", into = c("year", "month", "day"))
energy.upd$time <- format(as.POSIXct(strptime(energy.upd$Datetime,"%Y-%m-%d %H:%M",tz="")) ,format = "%H:%M")
library(dplyr)
energy.cln1 <- energy.upd%>%group_by(year, month)%>%summarise(consumption=mean(COMED_MW))
energy.cln2 <- energy.upd%>%group_by(year, month, day)%>%summarise(consumption=mean(COMED_MW))
energy.cln2$week <-  rep(1:52, each=7, length.out=2772)
energy.cln3 <- energy.cln2%>%group_by(year, week)%>%summarise(consumption=mean(consumption))
e.cln1.ts <- ts(energy.cln1$consumption, start = c(2011, 1), frequency = 12)
e.cln2.ts <- ts(energy.cln2$consumption, start = c(2011, 1), frequency = 365)
e.cln3.ts <- ts(energy.cln3$consumption, start = c(2011, 1), frequency = 52)
fix(energy.cln1)
energy.cln1$temp=c(42.8,    49.5,   61.3,   70.8,   72.8,   86.8,   91.4,   93.4,   80, 68.2,   57.9,   47.6,
                   50.4,    52.5,   64.3,   70.3,   77.9,   84.3,   87.7,   86.5,   80.0,   67.0,   59.7,   51.2,
                   49.1,    52.0,   56.4,   63.0,   72.3,   82.6,   84.5,   87.1,   82.4,   68.2,   53.5,   43.1,
                   45.3,    47.0,   55.1,   66.3,   74.4,   82.4,   83.8,   86.2,   80.3,   71.6,   51.5,   50.1,
                   44.5,    45.7,   56.1,   65.8,   70.9,   82.1,   87.1,   87.3,   82.7,   71.2,   58.7,   53.7,
                   47.0,    55.2,   61.2,   68.1,   72.5,   84.0,   87.4,   85.8,   81.5,   74.1,   63.5,   49.7,
                   51.2,    60.6,   65.7,   69.3,   75.4,   82.5,   86.6,   84.4,   80.6,   69.6,   62.4,   49.7,
                   45.8,    51.1,   63.3,   61.6,   79.0,   85.7,   88.8,   85.2)
energy.cln1$holidays = rep(c(2, 1, 0, 0, 1, 0, 1, 0, 1, 1, 2, 1), length.out=length(energy.cln1$consumption))
temp.cln1.ts <- ts(energy.cln1$temp, start = c(2011, 1), frequency = 12)
hd.cln1.ts <- ts(energy.cln1$holidays, start = c(2011, 1), frequency = 12)


########## Descriptive Analysis and Exploratory Analysis
par(mfrow=c(3, 1))
plot(e.cln2.ts, type = "o")
title("Hourly energy consumption by day")
plot(e.cln3.ts, type = "o")
title("Hourly energy consumption by week")
plot(e.cln1.ts, type = "o")
title("Hourly energy consumption by month")

############ training testing split
training <- energy.cln1[1:(round(0.8*nrow(energy.cln1))),]
testing <- energy.cln1[(round(0.8*nrow(energy.cln1))):nrow(energy.cln1) ,]

e.t.ts <- ts(training$consumption, start = c(2011, 1), frequency = 12)
e.testing.ts <- ts(testing$consumption, start = c(2011, 1), frequency = 12)
temp.t.ts <- ts(training$temp, start = c(2011, 1), frequency = 12)
```

```r
hd.t.ts <- ts(training$holidays, start = c(2011, 1), frequency = 12)

########## regression with autocorrelated errors
trend  = time(e.t.ts)
temp   = temp.t.ts - mean(temp.t.ts)
temp2  = temp^2
hd = hd.t.ts
summary(fit <- lm(e.t.ts~trend + temp + temp2+ hd, na.action=NULL))

#### add the autocorrelated errors
plot.ts(resid(fit))
acf2(resid(fit)) # possibly AR(2)
sarima(e.t.ts, 1,0,0, xreg=cbind(trend,temp,temp2, hd) ) # AIC 15.59315
sarima(e.t.ts, 2,0,0, xreg=cbind(trend,temp,temp2, hd) ) # AIC 15.56977 this one
reg1 <- sarima(e.t.ts, 2,0,0, xreg=cbind(trend,temp,temp2, hd) )

########## by month
plot(e.t.ts, type = "o")
title("Hourly energy consumption by month")
plot(stl(e.t.ts , s.window = "periodic"))# make decomposition of data.
le1 <- log(e.t.ts)
plot.ts(le1)
par(mfrow=c(4, 1))
plot(diff(le1, 1))
plot(diff(le1, 2))
plot(diff(le1, 3))
plot(diff(le1, 6))

dle1.0 <- diff(le1, 1)
acf2(dle1.0)
dle1 <- diff(le1, 3)
plot.ts(dle1)
acf2(dle1)
ddle1 <- diff(dle1, 12)
ddle1.0 <- diff(dle1.0,12)
par(mfrow=c(1,1))
plot.ts(ddle1)

par(mfrow=c(2,1))
monthplot(dle1)
monthplot(ddle1)

acf2(ddle1.0) # p,q 0 or 1; P, Q 0, 1 or 2
sarima(ddle1.0, 1,1,1, 0,1,1,12 ) # pass
sarima(ddle1.0, 1,1,1, 0,1,2,12 ) # aic -1.643136
sarima(ddle1.0, 0,1,1, 0,1,2,12 ) # aic -1.648339 this one
sarima(ddle1.0, 1,1,0, 0,1,2,12 ) # aic -1.405368
sarima(ddle1.0, 0,1,1, 0,1,1,12 ) # aic -1.610868
sarima(ddle1.0, 1,1,1, 0,1,1,12 ) # aic -1.632914
sarima(ddle1.0, 1,1,0, 0,1,1,12 ) # pass
sarima.for(e.t.ts, 18, 0,1,1, 0,1,2, 12 )
line(e.testing.ts)
plot(e.cln1.ts, type='o')

########## model validation and performance evaluation
library(forecast)
m1 <- sarima.for(e.t.ts, 18, 0,1,1, 0,1,2, 12 )
accuracy(m1$pred, testing$consumption[-1])


########## Spectral Analysis

mvspec(e.cln1.ts, spans=10, log='no')
mvspec(e.cln1.ts, log='no')
SigExtract(e.cln1.ts)
```