IE 360 Homework 2-3

1.The data has many perspectives to look at. To observe seasonality it can be tested from hourly, daily, weekly and monthly sums.

(All numbers shown in the first part are “additional” not “multiplicative”.)

In comparison between 24 hours of a day:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| AM Hours | -1.7439798 | -3.3398890 | -4.4525400 | -5.1737595 | -5.4904819 | -5.6302906 | -5.4583936 | -4.1165953 | -0.5083419 | 1.7543219 | 2.5255696 | 3.0688061 |
| PM Hours | 1.8428260 | 2.1673382 | 2.8209009 | 2.7051701 | 2.8754990 | 3.0202778 | 2.9816812 | 3.0656820 | 2.9514605 | 2.2929610 | 1.6469192 | 0.1948581 |

It can be seen that the biggest consume happens at the start of working hours and in the “prime time” of the day. Smallest consume happens at dawn.

With the same method, the days of the week can be compared.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
| Seasonality Factor | 20.988344 | 26.505490 | 29.382656 | 25.375071 | -9.774617 | -87.108119 | -5.368825 |

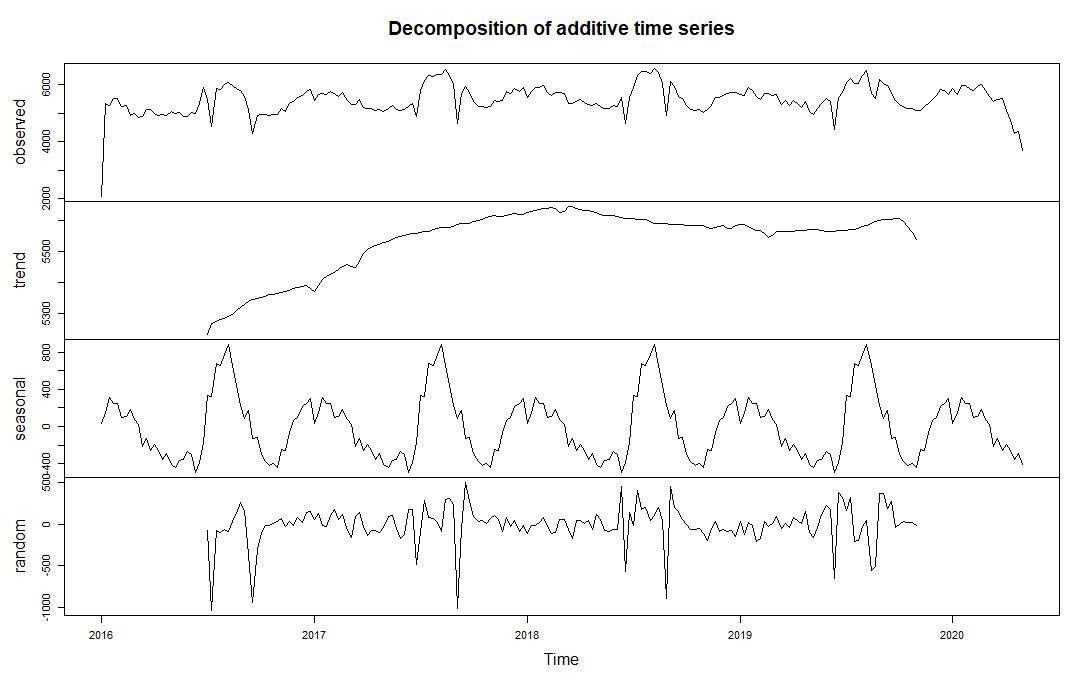
By the data, it is safe to say in weekends, consume is much smaller. This is natural since they are off days in most companies and factories. The reason for friday to be smaller than other weekdays in comparison, can be some machines and computers shutting down in the evening for coming weekend.

When it came to the months comparison:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
| 1595.5171 | -1455.8884 | -365.4215 | -1780.9136 | -896.9718 | -1377.9125 | 3155.5400 | 2577.8610 | -697.9590 | -1280.2736 | -633.3057 | 1159.7280 |

It’s harder to reach conclusions from this data since there is no obvious effect of seasons. And consecutive months have many ups and downs.

To have a better idea about trend and direction of consume it is beneficial to check decomposition of the data. Which is at the below based on weekly data.

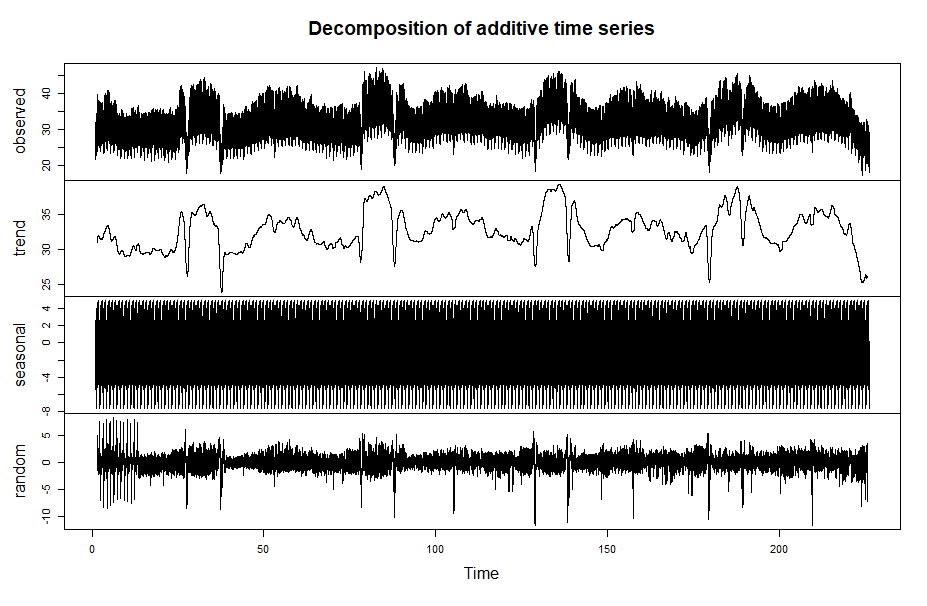


There is a clear trend of increase for the first two years and it is arguably stable since then. The first and last point of data being small is because they aren’t including whole week. However the drop in the last 4-5 weeks is probably due to half lock down process people are going thorugh. In addition to that seasonality of a year is a little bit more interpretable. Apparently, consume makes its peak on summer and goes on decreasing trend until winter. Which is a smaller peak probably caused by heating consume.

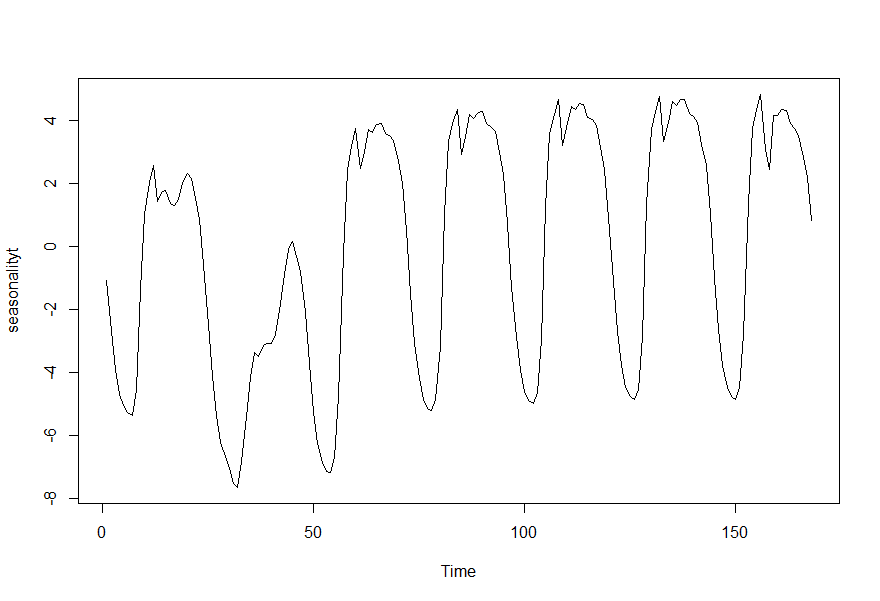
There is another part that can be observed. There are two huge drops every year with nearly same time of the year. They are due to religious holidays. At that times of the year nationwide off days occur. Which naturally stops working spaces to consume electricity.

2. To make further predictions and observations, it would be logical to decompose data from its seasonality and trend. To do that a period should be chosen for data to behave seasonal for that margin. Since nearly all of the working places work on weekdays, using the seasonality of this information is beneficial. To get a better result, this cycle can be combined with daily cycle which result a period with 7\*24=168 length.

Now at this point it shouldn’t be ignored that the data is too big to visualize every step with whole time margin(01.01.2016-26.04.2020).

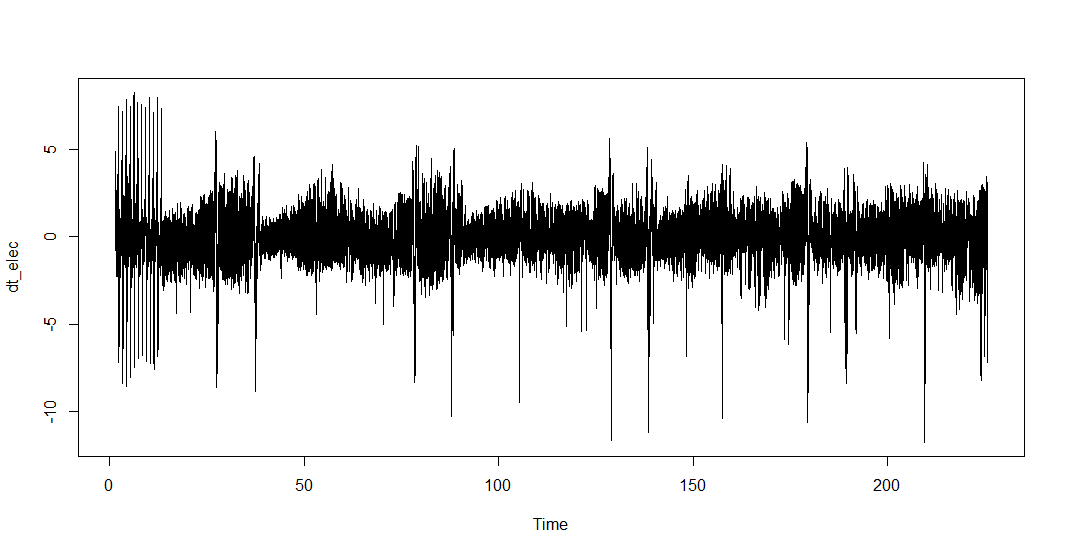


This is a table based on hourly data. The blackness in the 3rd row is supposed to show seasonality. The weekly cycle is below. And there are 225 of them in the 3rd row.



The data begins with the midnight of a friday and continues for a whole week. Even though the most of the days are similar(especially first four days of the week) it is useful for get the better result.

After taking seasonality and trend out of the data what left remain is random, as much as it can be.



3. To make a proper model the first step is the imitate the “random” above. In this assignment Auto-Regressive (AR) and Moving-Average(MA) is requested. arima function is useful for this step.

For creating an auto regressive model a lag should be chosen. And to test the goodness of lag AIC can be used. The standart AR (1,0,0) gives 63920.7 as AIC. To improve this score lag is increrased to 10. The downside of increasing the lag is it is really hard to calculate after a point. Even for computers. In addition to that the benefit of it gets smaller and smaller.

Coefficients:

ar1 ar2 ar3 ar4 ar5 ar6 ar7 ar8 ar9 ar10 intercept

1.4370 -0.6453 0.1636 -0.1108 0.0923 -0.0378 -0.0056 -0.0432 0.0222 0.0388 0.0023

s.e. 0.0051 0.0090 0.0096 0.0096 0.0097 0.0097 0.0096 0.0096 0.0090 0.0052 0.0279

sigma^2 estimated as 0.2324: log likelihood = -25970.62, aic = 51965.23

4.For MA model the standart is (0,0,1) and gives AIC= 95950.17 This result is worse than initial AR. Yet it can still be improved. For MA model window is open for changes. The result below is for (0,0,10)

Coefficients:

ma1 ma2 ma3 ma4 ma5 ma6 ma7 ma8 ma9 ma10 intercept

1.4419 1.446 1.3366 1.1371 0.9621 0.8102 0.6641 0.4683 0.2747 0.0685 0.0023

s.e. 0.0052 0.009 0.0113 0.0128 0.0140 0.0151 0.0154 0.0134 0.0093 0.0047 0.0237

sigma^2 estimated as 0.2301: log likelihood = -25787.07, aic = 51998.14

5. Since the AIC value of first model is smaller. It is used for prediction.

Time Series:

Start = c(226, 49)

End = c(226, 72)

Frequency = 168

[1] 0.001936947 0.001972049 0.002002963 0.002030439 0.002055104 0.002077448 0.002097825 0.002116468 0.002133521 0.002149071 0.002163178

[12] 0.002175896 0.002187290 0.002197442 0.002206453 0.002214436 0.002221511 0.002227796 0.002233401 0.002238423 0.002242942 0.002247024

[23] 0.002250720 0.002254067

The numbers above show the models’ prediction for the next day. To make a prediction of consume out of them, seasonality and trend should be added.

[49] -3.71062911 -5.24792560 -6.25827963 -6.88389006 -7.14136915 -7.19212193 -6.71568365 -4.77771719

[57] -0.22084527 2.40212534 3.16273700 3.74971717 2.46934296 3.01478901 3.71466552 3.63062786

[65] 3.87077699 3.91320661 3.59334475 3.52988949 3.37729388 2.68026579 1.98307792 0.49117561

The sum is:

Time Series:

Start = c(226, 49)

End = c(226, 72)

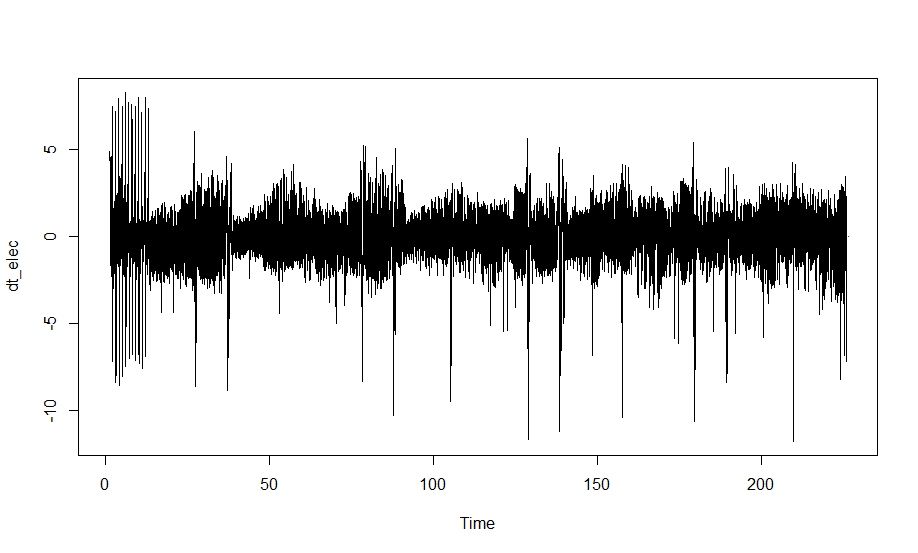
Frequency = 168

[1] -3.7086922 -5.2459536 -6.2562767 -6.8818596 -7.1393140 -7.1900445 -6.7135858 -4.7756007 -0.2187118

[10] 2.4042744 3.1649002 3.7518931 2.4715303 3.0169865 3.7168720 3.6328423 3.8729985 3.9154344

[19] 3.5955782 3.5321279 3.3795368 2.6825128 1.9853286 0.4934297

This data also requires trend. Which is 26.14387 when its latest calculation. So it needs to be added to all of the predictions above.



This little red dot is the prediction.