ASSIGNMENT 4

ADVANCED DATA ANALYSIS

Question 1: Two models are fit to the same data set. The results are given below.

Model 1:

```
                Estimate Std. Error t value Pr(>|t|)
Intercept       0.03818    0.20675   0.185    0.854
x               0.61672    0.12623   4.886   <0.001

R-Squared: 0.3321; Adjusted R-squared: 0.3182
F-statistic: 23.87 on 1 and 48 DF, p-value: < 0.001
```

Model 2:

```
                 Estimate Std. Error t value Pr(>|t|)
Intercept       -0.008506   0.208413  -0.041    0.968
x                0.360560   0.233740   1.543    0.130
x2                0.130633   0.100613   1.298    0.200

R-Squared: 0.3553; Adjusted R-squared: 0.3278
F-statistic: 12.95 on 2 and 47 DF, p-value: < 0.001
```

a- Which model would you select, Model 1 or 2? Why?

We choose Model 2 because it has relatively lower Std errors, smaller t.values as well as better R^2 and adjusted R^2.

b- Write the equation of the selected model

$$\boldsymbol{Y = 0.360560\ X + 0.130633\ X^2 - 0.008506}$$

Question 2: In dataset with wage, education and age features, where the main task is to find the best model to predict wage based on education level and age, we designed 4 models. Then anova test is used to compare these models.

```
> anova(Model1,Model2,Model3,Model4)
Analysis of Variance Table
Model 1: wage ~ education
Model 2: wage ~ education + age
Model 3: wage ~ education + poly(age, 2)
Model 4: wage ~ education + poly(age, 3)
  Res.Df     RSS Df Sum of Sq        F   Pr(>F)
1   2995 3995721
2   2994 3867992  1    127729 102.7378  <2e-16 ***
3   2993 3725395  1    142597 114.6969  <2e-16 ***
4   2992 3719809  1      5587   4.4936  0.0341 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Which model is the best to use? and why?

All the four models have significant p.values (<0.05) where RSS decreases from Model 1 till Model 4. Meaning model 4 is the best performing one. Taken together, the 4th Model (with education, and cubic age) is the one we choose (its p.value is bigger than the others however still significant less than 5%, smaller RSS)

Question 3: A curve is fit with basis functions

$$y = \begin{cases} \boldsymbol{b1(X) = \beta_1}\,(X+1) & \boldsymbol{for\ 0 \le X \le 2} \\ \boldsymbol{b2(X) = \beta_2(X^2 - 8X)} & \boldsymbol{for\ 2 < X \le 6} \\ \boldsymbol{b3(X) = \beta_3}\,(X-9) & \boldsymbol{for}\ 6 < X \le 8 \end{cases}$$

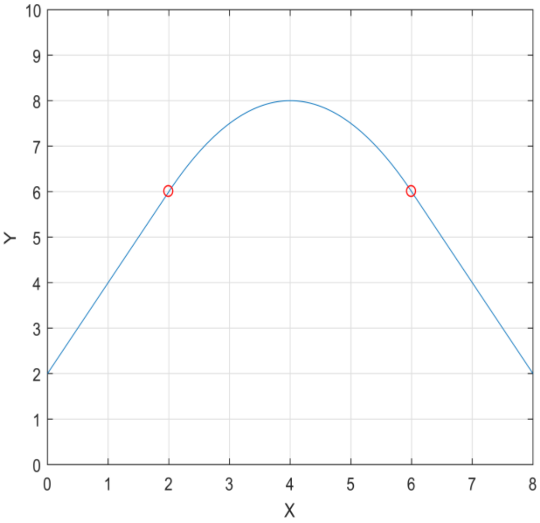
And corresponding coefficients β1, β2, and β3, where β1 = 2 and β3 = −2.
What coefficient β2 converts this curve into a quadratic spline?

$$\boldsymbol{b2(X) = \beta_2(X^2 - 8X)\ for\ 2 < X \le 6}$$

$$\boldsymbol{\beta_2 = \frac{b2(X)}{(X^2 - 8X)}\ ;where\ 2 < X \le 6}$$

We calculate $\boldsymbol{\beta_2}$ for $\boldsymbol{X = 6}$ for example
$\boldsymbol{b2(X = 6) = 6}$

➔ $\boldsymbol{\beta_2 = \frac{-1}{2}}$



Question 4: What are the types of missing in data? (At least three types) Explain one suitable solution to handle each of these types of missing.

1. **MCAR missing completely at random, we can here use Listwise deletion if the dataset is large enough and/or the NAs are only a few.**
2. **Missing at random (MAR), we can use Maximum Likelihood Estimation.**
3. **Missing not at random (MNAR), we can use Dummy variable adjustment.**

Question 5: What is the "multiple imputations"? How do we get the analysis results after the multiple imputations?

Multiple imputation is a general approach to the problem of missing data that is available in several commonly used statistical packages. It aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets (for example by using mean, regression, maximum likelihood…etc) and appropriately combining results obtained from each of them. We average the values of the parameter estimates across the samples to produce a single point estimate.

Question 6: What is the t-test, and what are the main assumptions to using the t-test?

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. Calculating a t-

test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group. Essentially, a t-test allows us to compare the average values of the two data sets and determine if they came from the same population. There are several different types of t-test that can be performed depending on the data and type of analysis required.

Question 7: the following R Command is executed, and we got the following output where LungCap is the lung capacity and Gender is male or female

```
> t.test(LungCap ~ Gender,  data = dataset,  var.equal = TRUE, paired = FALSE)

data:  LungCap by Gender
t = -4.6336, df = 723, p-value = 4.262e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.2864335 -0.5207397
sample estimates:
mean in group female   mean in group male
            7.405746             8.309332
```

What is the t-test type applied in the command? Based on the output above, what is the conclusion from this test?

Paired t.test was used here and we can see that the test is significant (small p.value). The conclusion we can get from this analysis that Males have larger lung capacity than females (we can even see the mean for each group where it's 7.5 for females and 8.30 for males).

Question 8: answer the following about the ANOVA test:

a- What does one-way or two-way ANOVA means?

A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA.

b- What does ANOVA Tell us?

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

c- What are the types of one-way ANOVA?

MANOVA, Factorial ANOVA

Question 9: In the following, we apply one-way ANOVA on data of weight loss of subjects when they use one of three diet programs. ANOVA test and Post hoc test are shown below. **Report the results of this output**

```
> output = aov(weight.loss ~ Diet, data = diet1)
> summary(output)
             Df Sum Sq  Mean Sq   F value    Pr(>F)
```

```
Diet          1   45.8   45.78     7.639      0.00716 **
Residuals    76  455.5    5.99
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> HSD.test(output, "Diet", group = FALSE, console = TRUE)

Study: output ~ "Diet"
HSD Test for weight.loss
Mean Square Error:  5.993315
Diet,  means

  weight.loss      std  r  Min Max
1    3.300000 2.240148 24 -0.6 9.0
2    3.025926 2.523367 27 -2.1 7.9
3    5.148148 2.395568 27  0.5 9.2

Alpha: 0.05 ; DF Error: 76
Critical Value of Studentized Range: 3.380649

Comparison between treatments means

      difference pvalue signif.       LCL        UCL
1 - 2  0.2740741 0.9161         -1.367711  1.9158587
1 - 3 -1.8481481 0.0235       * -3.489933 -0.2063635
2 - 3 -2.1222222 0.0059      ** -3.714987 -0.5294572
```

As the p-value from “aov” is less than the significance level 0.05, we can conclude that there are significant differences between the diets. In one-way ANOVA test, a significant p-value indicates that some of the diets’ means are different, but we don’t know which pairs of groups are different and that’s why we perform the second piece of code.
We can see that Diet 1 is the one that offers the highest weight loss, followed by Diet 1 and finally Diet 2. Also, when comparing each pair, we can see that Diet 1 and Diet 2 are not significantly different (p.value=0.9161) while Diets 1 & 2 and Diets 1 &3 are significantly different from one another.

Question 10: Choose any time series dataset from here or other dataset resources, and conduct an analysis on the dataset as you did in assignment 3. Place the analysis without the code (you can attach the code files or notebook files with the assignment).

For this question, I will attach the used dataset that I have chosen to work on.

Report on the study of the time series of electrical energy consumption in Tunisia

I/- Description of the database:

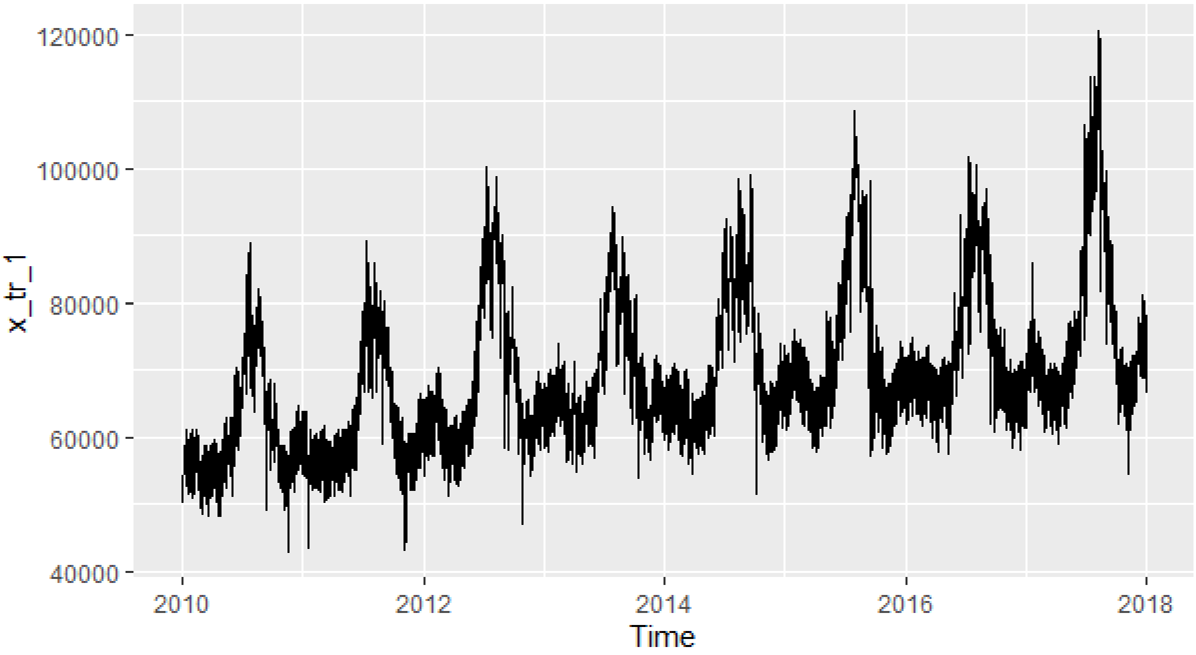
* The "electricity" database presents the volume of daily energy consumption in Tunisia between the years 2010 and 2017. It also contains the daily temperatures (min, max and average), the days (from Monday to Sunday), public holidays, days belonging to the month of Ramadan and the different months.

II/-Study:

*We will try to study the series, break it down, model it (by focusing on the variables that explain an increase in energy consumption which are a priori the temperature, public holidays and the days of the month of Ramadan) and finally predict the level of future energy consumption.

1- Visualization:

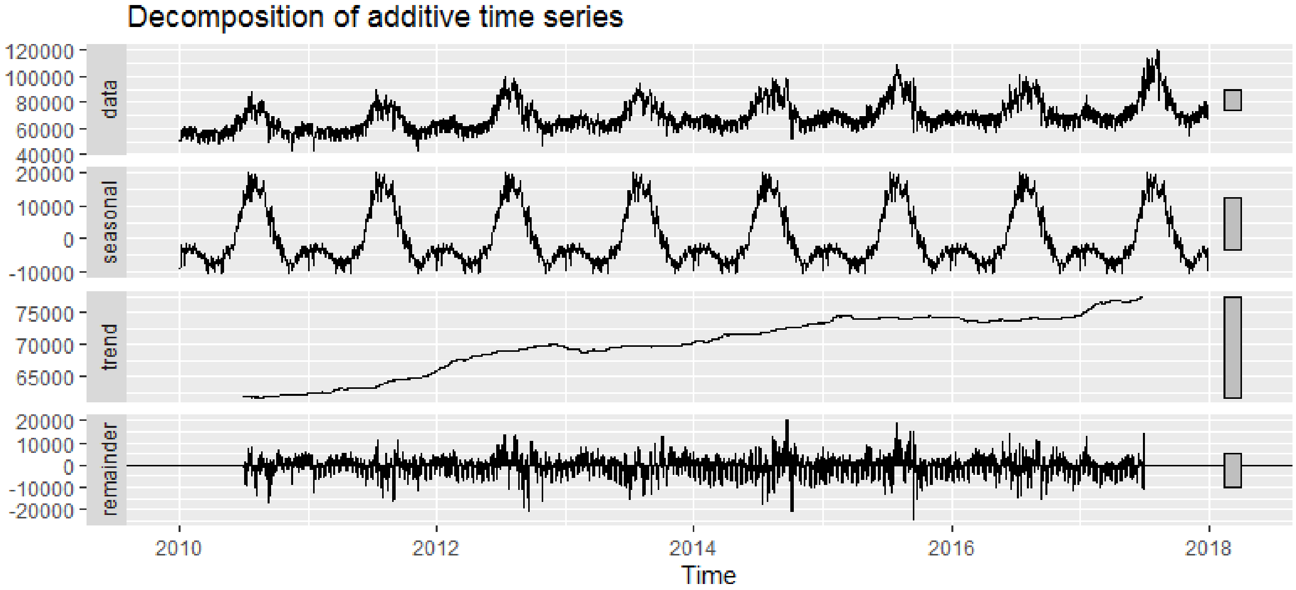
*Let's start with the visualization of the series:



We can notice a priori the presence of an annual seasonality (365 days) as well as the presence of a trend.

2- Decomposition:

To confirm these remarks, we move on to the decomposition of the series:



After decomposition, we can clearly visualize the seasonality and the trend of the series as well as the noise.

Note: Using the decompose() function on R one can obtain the result presented below. It should also be noted that R, according to the series passed as a parameter, chooses the adequate mode of decomposition which is in our case “additive”.

3- Prediction: using stlm and stlf

We will start with the prediction of the series without taking into account only the history of the energy consumption: We use the stlf function of R where the predictions of STL objects are obtained by applying a non-seasonal prediction method to the seasonally adjusted data and further seasonal adjustment using the last year of the seasonal component.

stlf combines stlm and forecast. stlm. It takes a ts argument, applies an STL decomposition, models the seasonally adjusted data, re-seasonalizes, and returns the forecasts. However, it allows more general forecasting methods to be specified via forecast function.

We get the following result:

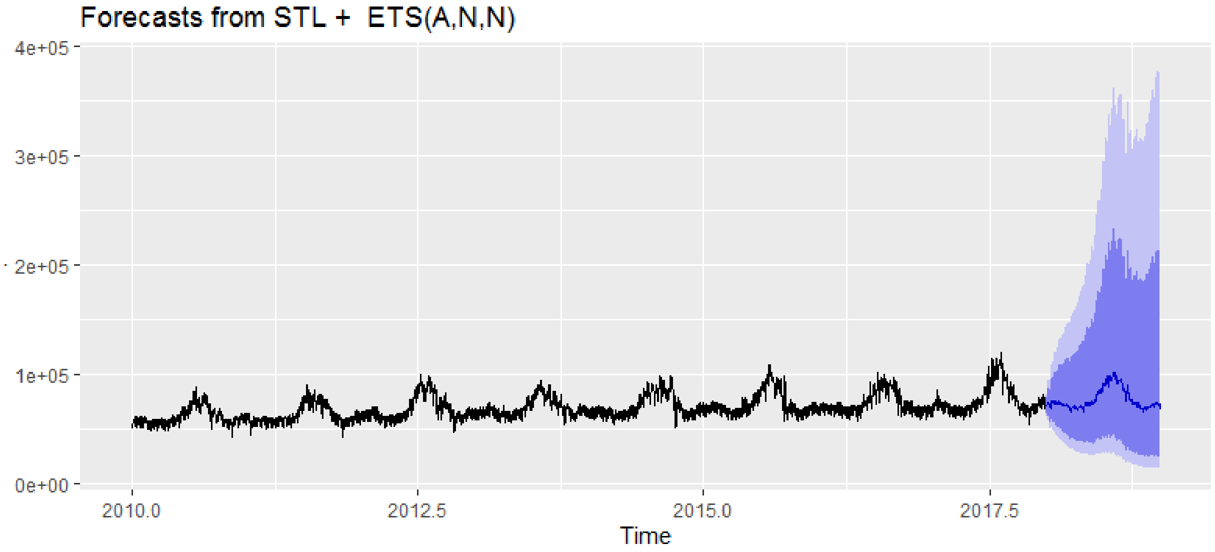


Figure 1: Forecast using stlf

To improve the prediction quality an uses the tslm function of R which is used to fit linear models to time series, including trend and seasonality components.

The tslm function is designed to fit linear models to time series data. It is intended to approximately mimic lm (and calls lm to do the estimation), but to package the output to remember the ts attributes. It also handles some predictor variables automatically, notably trend and season.

We get the following result:

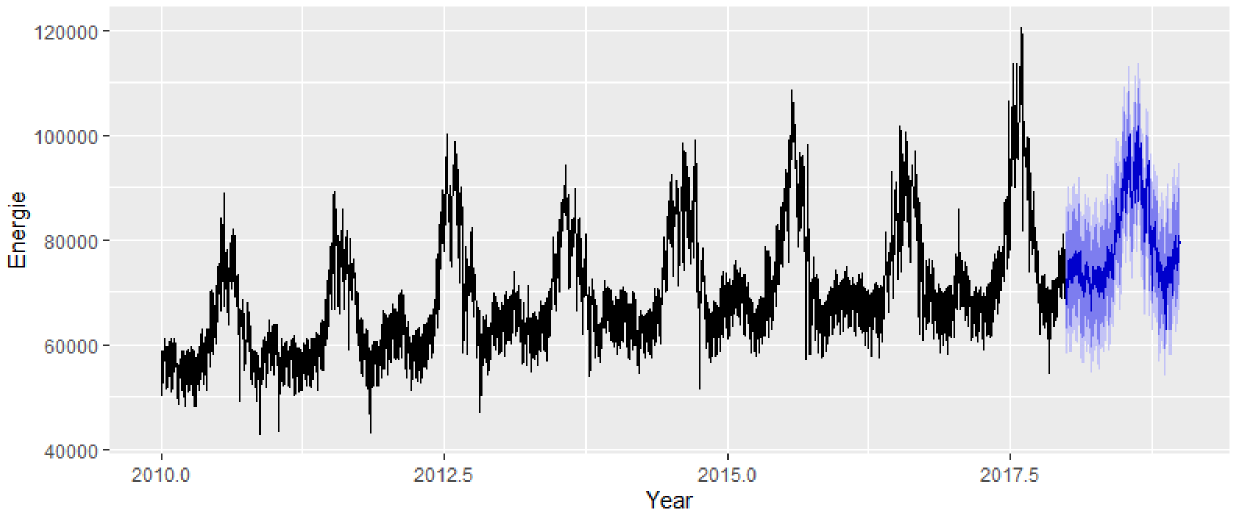


Figure 2: Prediction using tslm

We then notice that the forecast which takes into account the seasonality and the trend of the series is better.

*Now we will take into consideration the other variables of the base while trying to make the prediction of the time series:

3-1- Taking into account weekends, public holidays, Ramadan days and the average temperature:

In this part we will predict the series by taking the average temperature and playing on the days (holidays or not, weekends or not and belonging to Ramadan or not).

We will extract a prediction of the extreme cases: maximum possible energy consumption and minimum possible energy consumption.

The result is the following:

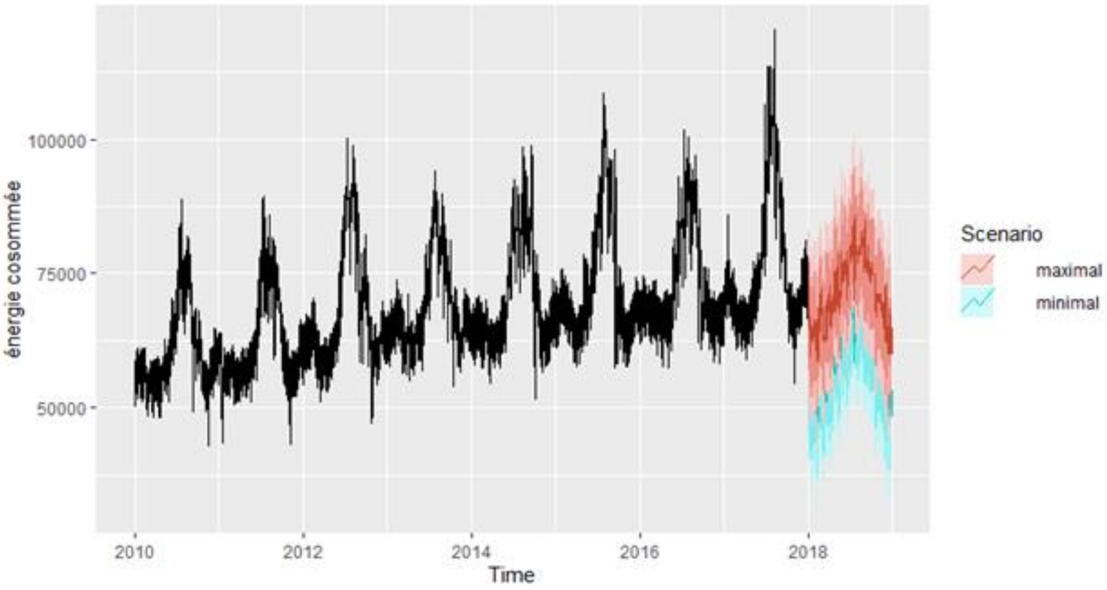


Figure 3:Max & Min Prediction using Temperature and holidays

Using the graph below, we can notice that the energy consumption levels are a little more compared to those obtained before (in figure 4), while keeping the same trend. This result can be explained by the fact that other variables were implemented in the model used for the estimation.

→ To better discern the variables that explain the most the 'energy consumption' variable, we will now reconstruct the model by involving only the temperature variable, which is logically the variable that has the greatest effect on our energy needs (air conditioning, refrigeration, etc.).

3-2- Taking into account the temperature variable:

In this part also we will take the two extreme cases, that is to say, we will model and predict the series according to the maximum temperature and the minimum temperature.

The result obtained is as follows:

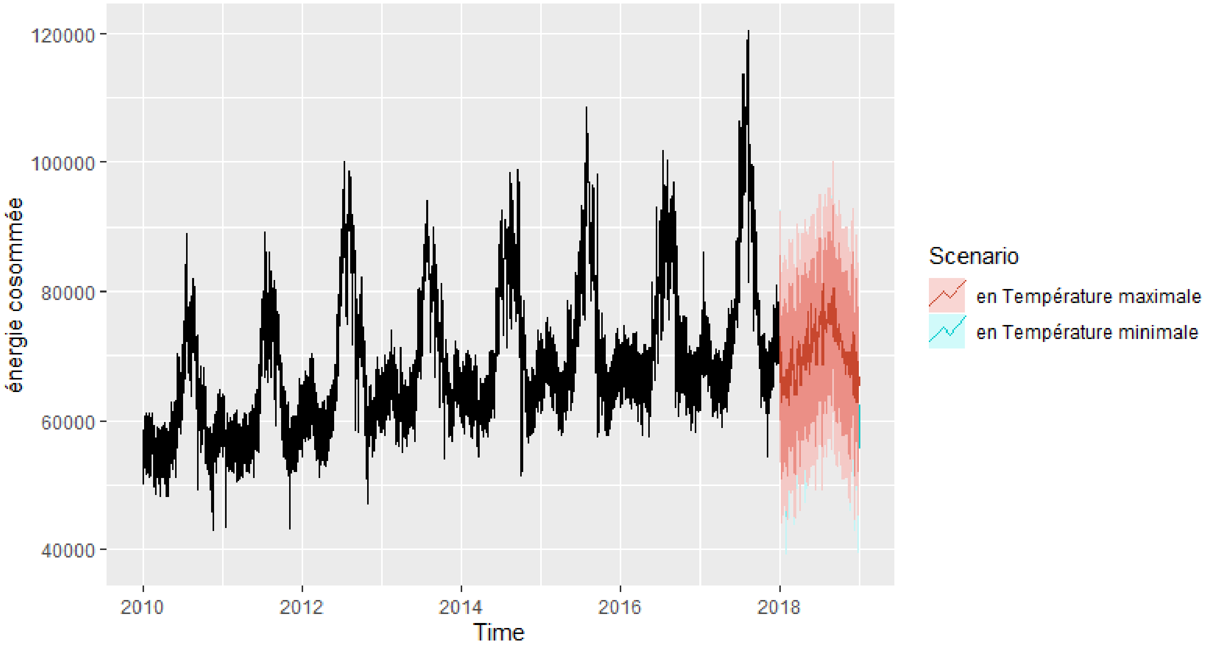


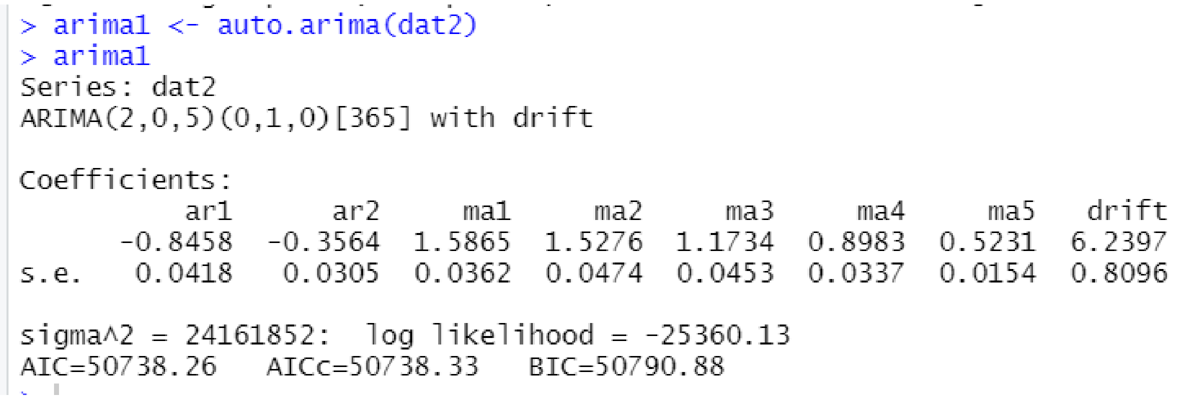
Figure 4: Max & Min Prediction using Temperature

The graph below reflects the expected result, which indicates that temperatures, and more precisely high temperatures, affect the increase in energy consumption the most.

Note: It can be explained that the fact that the minimum temperature part of the scenario is almost hidden is due to the existence of similar temperature values between Tmin.

4- Univariate time series analysis:

We can run the auto-arima function in R to find the most accurate model:



```
> arima1 <- auto.arima(dat2)
> arima1
Series: dat2
ARIMA(2,0,5)(0,1,0)[365] with drift

Coefficients:
          ar1      ar2     ma1     ma2     ma3     ma4     ma5   drift
      -0.8458  -0.3564  1.5865  1.5276  1.1734  0.8983  0.5231  6.2397
s.e.   0.0418   0.0305  0.0362  0.0474  0.0453  0.0337  0.0154  0.8096

sigma^2 = 24161852:  log likelihood = -25360.13
AIC=50738.26   AICc=50738.33   BIC=50790.88
```

As seen above the best model found is **ARIMA(2,0,5)(0,1,0)[365]** with drift.
We use this model to predict the next year's energy consumption:

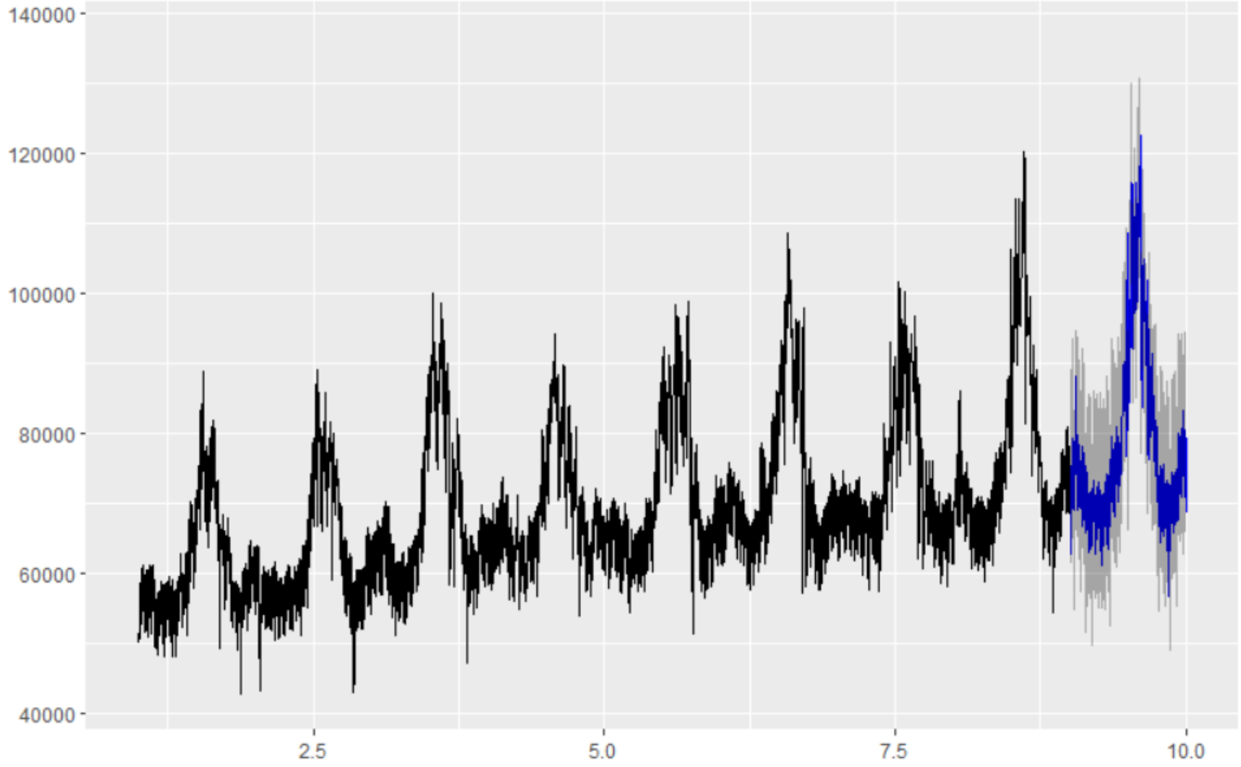


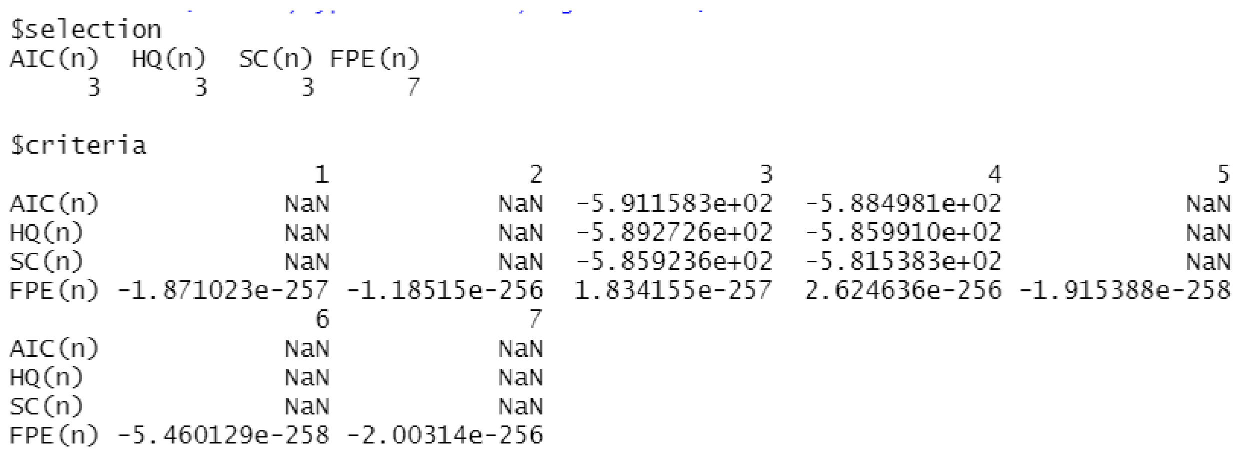
Figure 5: Prediction using ARIMA(2,0,5)(0,1,0)[365]

As seen in Figure 5 the ARIMA(2,0,5)(0,1,0)[365] Model gives quite decent prediction especially compared to the ones obtained by tslm (Figure 2).

The final ARIMA model helps us well-predict the electrical energy consumption in Tunisia.

4- Multivariate time series analysis:

Now we move on to forming one Multivariate time series that has all the possible univariate time series we have.
We first test whether these times series are stationary or not, we can use adf test.
We find that except Temperature related ones (Min, Max, Moy (=mean)) are stationary.
To make the time series stationary we of course differentiate.
Now that out time series is ready, we can start VAR modelling.
To do so, we need to find the good parameter for our VAR modelling. One way to do so is through using the "VARselect" algorithm:

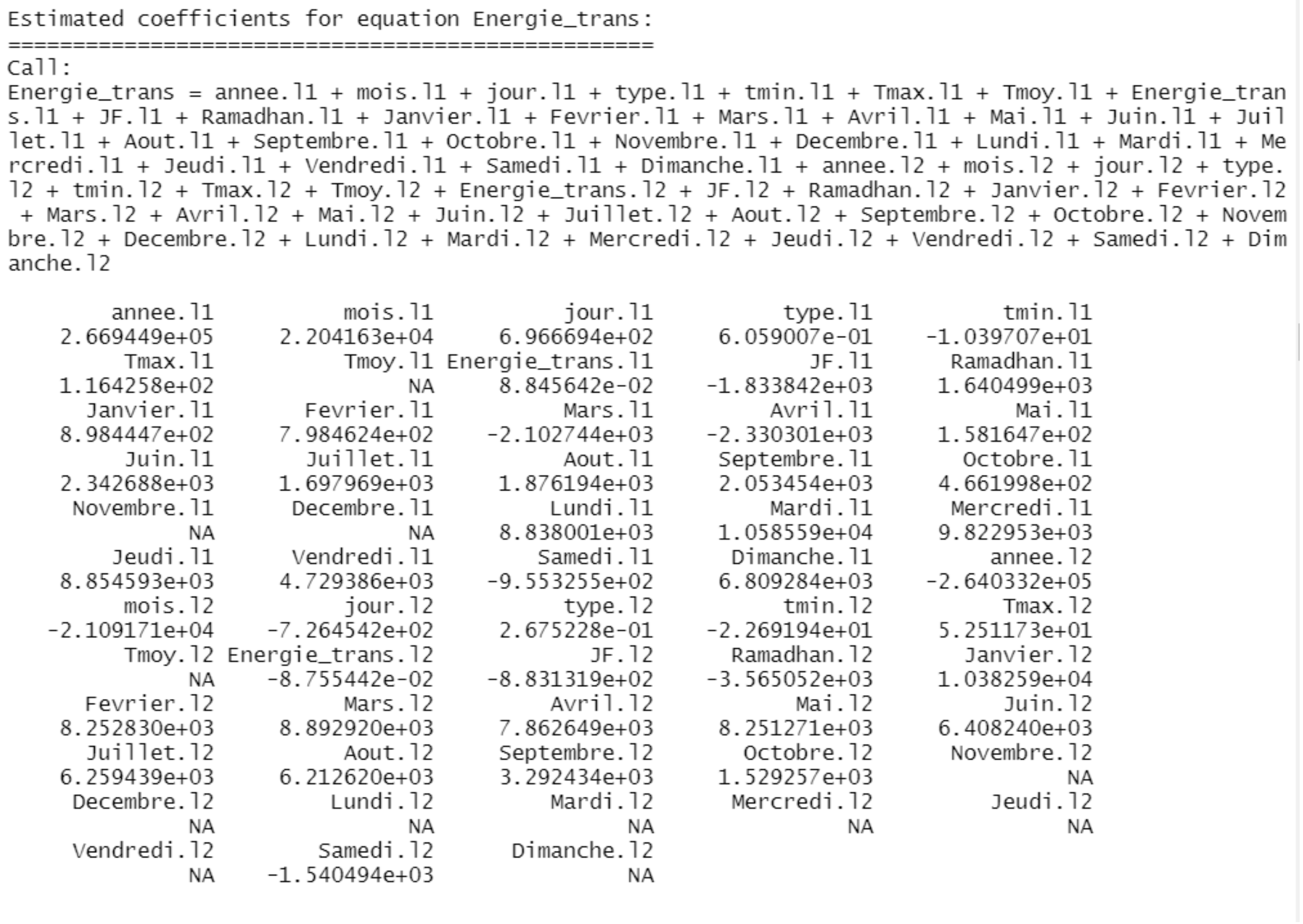


```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n) 
     3      3      3      7 

$criteria
                    1             2             3             4             5
AIC(n)            NaN           NaN  -5.911583e+02  -5.884981e+02           NaN
HQ(n)             NaN           NaN  -5.892726e+02  -5.859910e+02           NaN
SC(n)             NaN           NaN  -5.859236e+02  -5.815383e+02           NaN
FPE(n) -1.871023e-257 -1.18515e-256  1.834155e-257  2.624636e-256 -1.915388e-258
                    6             7
AIC(n)            NaN           NaN
HQ(n)             NaN           NaN
SC(n)             NaN           NaN
FPE(n) -5.460129e-258 -2.00314e-256
```

Observing the selection algorithm results we can conclude that the lag.max=3 model is the optimal option as the selection criteria (AIC, HQ, SC) are better for this one.

We can now create the VAR model with information obtained from the prior selection. The results show the possible models for each one of variables. As we are only interested in the variable "Energy consumption", we can focus only on that part seen bellow:



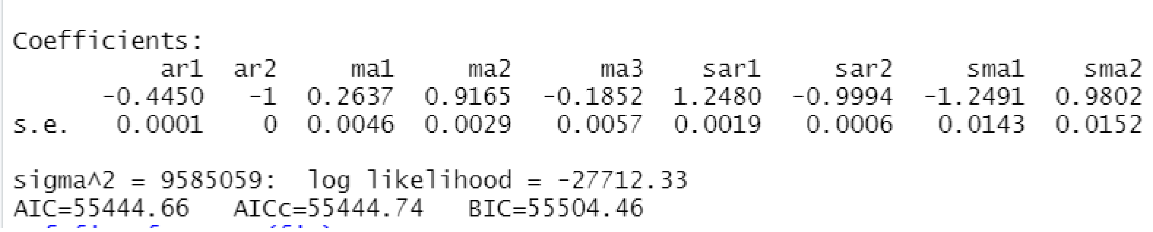
```
Estimated coefficients for equation Energie_trans:
==================================================
Call:
Energie_trans = annee.l1 + mois.l1 + jour.l1 + type.l1 + tmin.l1 + Tmax.l1 + Tmoy.l1 + Energie_tran
s.l1 + JF.l1 + Ramadhan.l1 + Janvier.l1 + Fevrier.l1 + Mars.l1 + Avril.l1 + Mai.l1 + Juin.l1 + Juil
let.l1 + Aout.l1 + Septembre.l1 + Octobre.l1 + Novembre.l1 + Decembre.l1 + Lundi.l1 + Mardi.l1 + Me
rcredi.l1 + Jeudi.l1 + Vendredi.l1 + Samedi.l1 + Dimanche.l1 + annee.l2 + mois.l2 + jour.l2 + type.
l2 + tmin.l2 + Tmax.l2 + Tmoy.l2 + Energie_trans.l2 + JF.l2 + Ramadhan.l2 + Janvier.l2 + Fevrier.l2
 + Mars.l2 + Avril.l2 + Mai.l2 + Juin.l2 + Juillet.l2 + Aout.l2 + Septembre.l2 + Octobre.l2 + Novem
bre.l2 + Decembre.l2 + Lundi.l2 + Mardi.l2 + Mercredi.l2 + Jeudi.l2 + Vendredi.l2 + Samedi.l2 + Dim
anche.l2

       annee.l1         mois.l1         jour.l1         type.l1         tmin.l1 
   2.669449e+05    2.204163e+04    6.966694e+02    6.059007e-01   -1.039707e+01 
        Tmax.l1         Tmoy.l1 Energie_trans.l1           JF.l1     Ramadhan.l1 
   1.164258e+02              NA    8.845642e-02   -1.833842e+03    1.640499e+03 
     Janvier.l1      Fevrier.l1         Mars.l1         Avril.l1          Mai.l1 
   8.984447e+02    7.984624e+02   -2.102744e+03   -2.330301e+03    1.581647e+02 
        Juin.l1      Juillet.l1         Aout.l1    Septembre.l1      Octobre.l1 
   2.342688e+03    1.697969e+03    1.876194e+03    2.053454e+03    4.661998e+02 
    Novembre.l1     Decembre.l1        Lundi.l1        Mardi.l1     Mercredi.l1 
             NA              NA    8.838001e+03    1.058559e+04    9.822953e+03 
       Jeudi.l1     Vendredi.l1       Samedi.l1     Dimanche.l1        annee.l2 
   8.854593e+03    4.729386e+03   -9.553255e+02    6.809284e+03   -2.640332e+05 
        mois.l2         jour.l2         type.l2         tmin.l2         Tmax.l2 
  -2.109171e+04   -7.264542e+02    2.675228e-01   -2.269194e+01    5.251173e+01 
        Tmoy.l2 Energie_trans.l2           JF.l2     Ramadhan.l2      Janvier.l2 
             NA   -8.755442e-02   -8.831319e+02   -3.565052e+03    1.038259e+04 
     Fevrier.l2         Mars.l2        Avril.l2          Mai.l2         Juin.l2 
   8.252830e+03    8.892920e+03    7.862649e+03    8.251271e+03    6.408240e+03 
     Juillet.l2         Aout.l2    Septembre.l2      Octobre.l2     Novembre.l2 
   6.259439e+03    6.212620e+03    3.292434e+03    1.529257e+03              NA 
    Decembre.l2        Lundi.l2        Mardi.l2     Mercredi.l2        Jeudi.l2 
             NA              NA              NA              NA              NA 
    Vendredi.l2       Samedi.l2     Dimanche.l2 
             NA   -1.540494e+03              NA 
```

As seen above the variables that are significant are:

- Energie_trans.l2 : Energy consumption with lag=2
- Energie_trans.l1 : Energy consumption with lag=1
- type.l1 : Type with lag=1
- type.l2 Type with lag=2

Observing the performance of the model obtained from the VAR model seen bellow we can see that based on the scores (AIC, BIC) this model performs worse than the one obtained in the univariate one.



```
Coefficients:
          ar1  ar2     ma1     ma2      ma3    sar1     sar2     sma1    sma2
      -0.4450   -1  0.2637  0.9165  -0.1852  1.2480  -0.9994  -1.2491  0.9802
s.e.   0.0001    0  0.0046  0.0029   0.0057  0.0019   0.0006   0.0143  0.0152

sigma^2 = 9585059:  log likelihood = -27712.33
AIC=55444.66   AICc=55444.74   BIC=55504.46
```

➔ In conclusion, we choose the model **ARIMA(2,0,5)(0,1,0)[365]** as the best one for predicting the electrical energy consumption in Tunisia.