uc3m | Universidad Carlos III de Madrid

Master in Big Data Analytics
2017/2018

Master's Thesis

France Regional Electricity Consumption Clustering Using Generalised Cross Correlation.

Pierre Mercatoris

Supervisors
Daniel Peña Sánchez de Rivera
Andrés Modesto Alonso Fernández
Madrid y fecha de presentación prevista

Keywords: Clustering, time series, electricity consumption, distance metric.
Summary: This work attempts to characterise the electricity consumption of the regions of France between 2013 and late 2017. It does this by applying the Generalised Cross Correlation allowing the clustering of time series by their linear dependency. Each cluster's trend and consumption patterns are then discussed.

Contents

List of Figures

1 TODO Introduction[0/2]

1.1 TODO Cluster electricity consumption using GCC

[Jain and Dubes, 1988]

1.2 TODO Clustering time series

2 TODO Methodology[3/4]

2.1 DONE Data description

The electricity consumption was available at a 30 minutes frequency for each of the 12 regions of France from 2013 to 2017. Each year of each region can be downloaded from the French transmission operator (Rte) download portal[1].

Consumption from January 2013 to September of 2017 were downloaded for each of the 12 metropolitan mainland regions of France (excluding Corsica).

However, those regions are still very young, as before 2016, those were 21 separate regions. Regions in France lack separate legislative power, but can manage a considerable part of their budget for main infrastructures such as education, public transport, universities and research, and help to businesses. It is therefore expected to find some interesting clusters, where we might see some reminiscence of the old regions.

2.2 DONE Data preparation

2.2.1 DONE Cleaning

The complete data set was spread across 60 different tables (years and regions) that were merged into one large table (table 1).

Table 1: Original data structure.

Périmètre	Date	Heures	Consommation
Grand-Est	2016-01-01	00:00	5130
Grand-Est	2016-01-01	00:15	
Grand-Est	2016-01-01	00:30	5130
Grand-Est	2016-01-01	00:45	
Grand-Est	2016-01-01	01:00	5014
.....			

As data rarely comes clean, there were some imperfections in the names of the regions. Some days the regions were named after the old ones e.g. Languedoc-Roussillon et Midi-Pyrénées instead of Occitanie, or Aquitaine, Limousin et Poitou-Charentes instead of Nouvelle-Aquitaine.

With the raw data cleaned from imperfections, each column was formatted to required data type. A pivot table was then used so as to move each region as a column, and each row is a consumption measurement. The date then needed to be set as UTC in order to avoid problems at the summer/winter time change. As the original frequency of the data is 15 minutes but as there are only data every 30 minutes, the table was resampled by taking the sum for each 30 minutes, resulting in the table below (table 2).

The region with the highest consumption are observed in the Iles-de-France and the lowest in the Centre-Val de Loire. We can also clearly see yearly seasonality with higher consumption during winter times (figure 1).

The pivot table was used again so that each time of the day is a columns, and each row is a daily value for a certain time and region, the resulting table has 576 columns (48 x 12 regions) and 1794 rows/days.(table 3).

[1] `http://www.rte-france.com/en/eco2mix/eco2mix-telechargement-en`

Table 2: Regional series before splitting the series by time of the day.

Périmètre	Auvergne-Rhône-Alpes	Bourgogne-Franche-Comté	...
Datetime			
2013-01-01_00:00:00+00:00	NaN	NaN	...
2013-01-01_00:30:00+00:00	8173.0	2357.0	...
2013-01-01_01:00:00+00:00	7944.0	2289.0	...
2013-01-01_01:30:00+00:00	7896.0	2326.0	
2013-01-01_02:00:00+00:00	7882.0	2409.0	

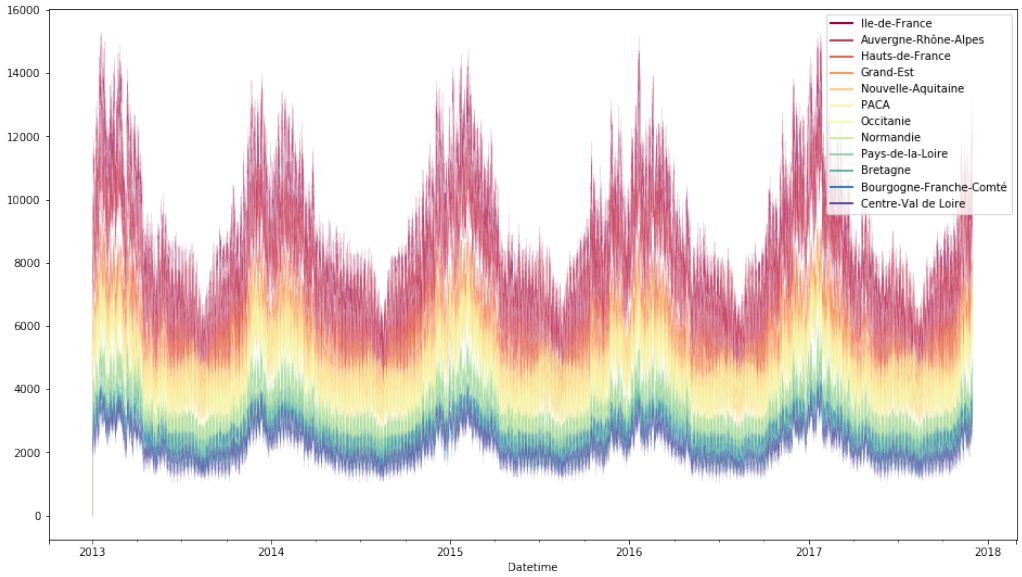

Figure 1: Mean electricity consumption of each of the french regions from 2013 to end 2017.

Table 3: Final data format before export to csv.

Périmètre	Auvergne-Rhône-Alpes		
time	00:00:00	00:30:00	01:00:00
date			
2013-01-02	7847.0	7674.0	7427.0
2013-01-03	9028.0	8839.0	8544.0
2013-01-04	8982.0	8754.0	8476.0
2013-01-05	8625.0	8465.0	8165.0
2013-01-06	8314.0	8097.0	7814.0

In figure 2, we can already see that consumption midday is much higher than at night, with more spread in the summmer than in the winter.

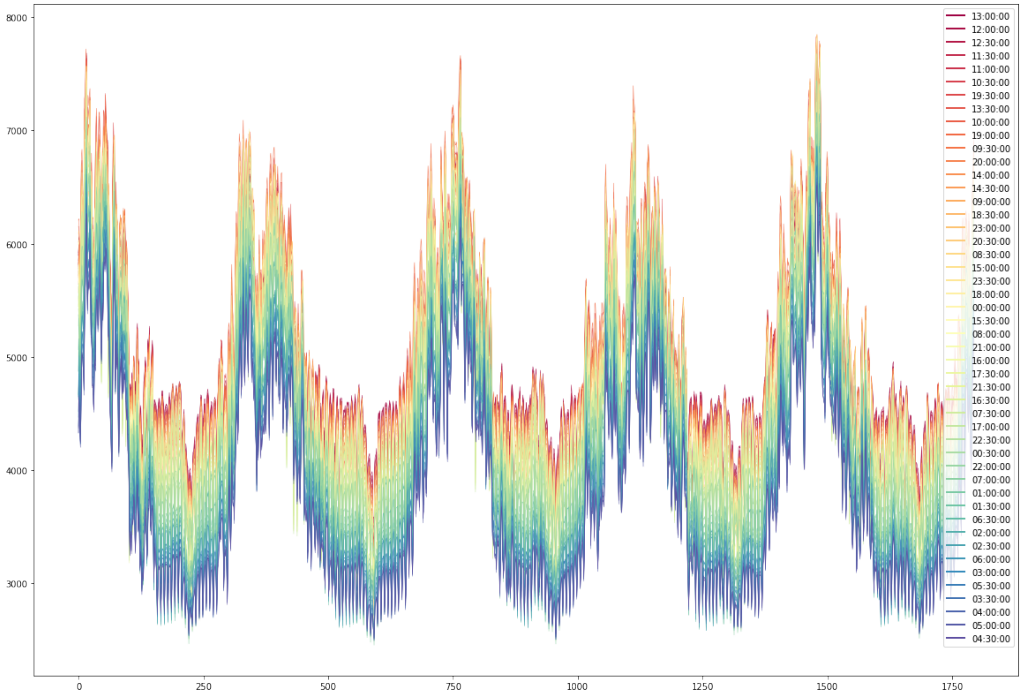

Figure 2: Mean electricity consumption for all the regions of France at different times.

2.2.2 DONE Transformation

1. **DONE** Stationarity The original series feature a strong seasonality as show in figure 3.

   To try and remove it, the weekly difference was taken (difference between all the values separated by 7 days). This was able to remove most of the seasonality (fig. 4).

   So as to get as close to stationarity as possible without loosing too much data, another difference was tken, but this time only 1 day. Now, most of the values stay within the confidence interval (fig. 5).

   The Dickey-Fuller test was used on all the series and confirmed that all the series are now significantly stationary (all p-values lower than $10e^{-21}$).

2. **DONE** Standardisation In order to standardise the data so as to get a mean of 0 and standard deviation of 1, the z-score was applied to each individual series (1).

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

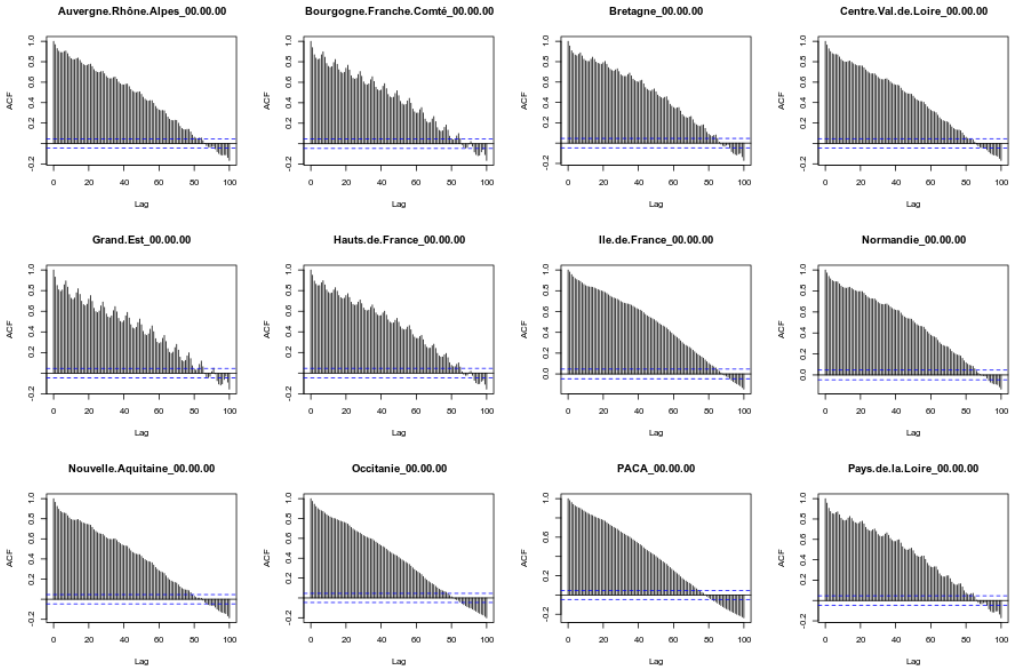

Figure 3: Autocorrelation function of the original data.

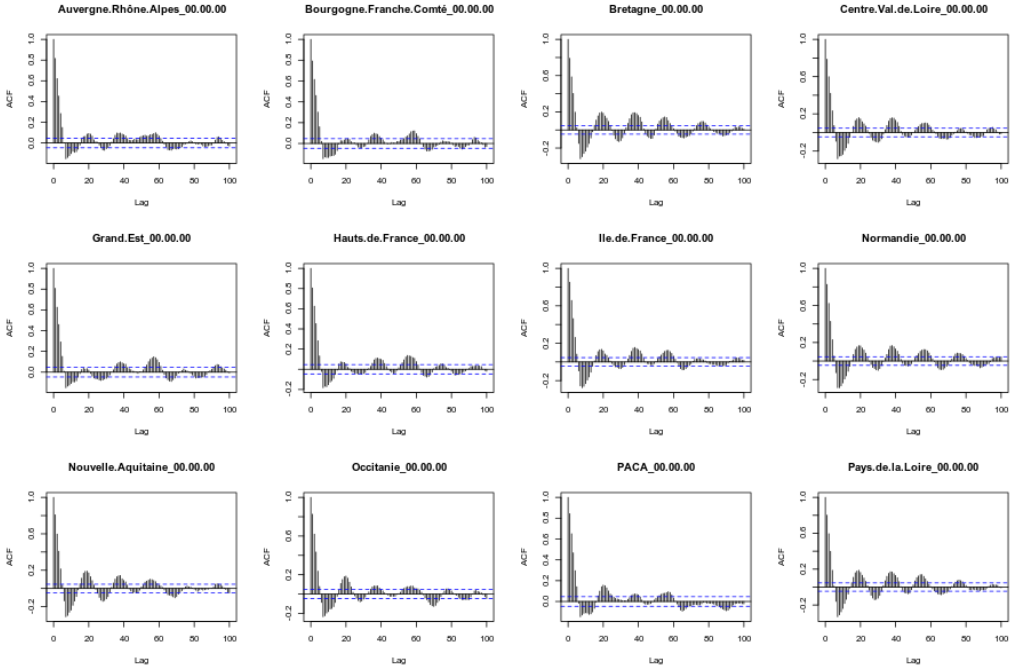

Figure 4: Autocorrelation function of the weekly differentiated series.

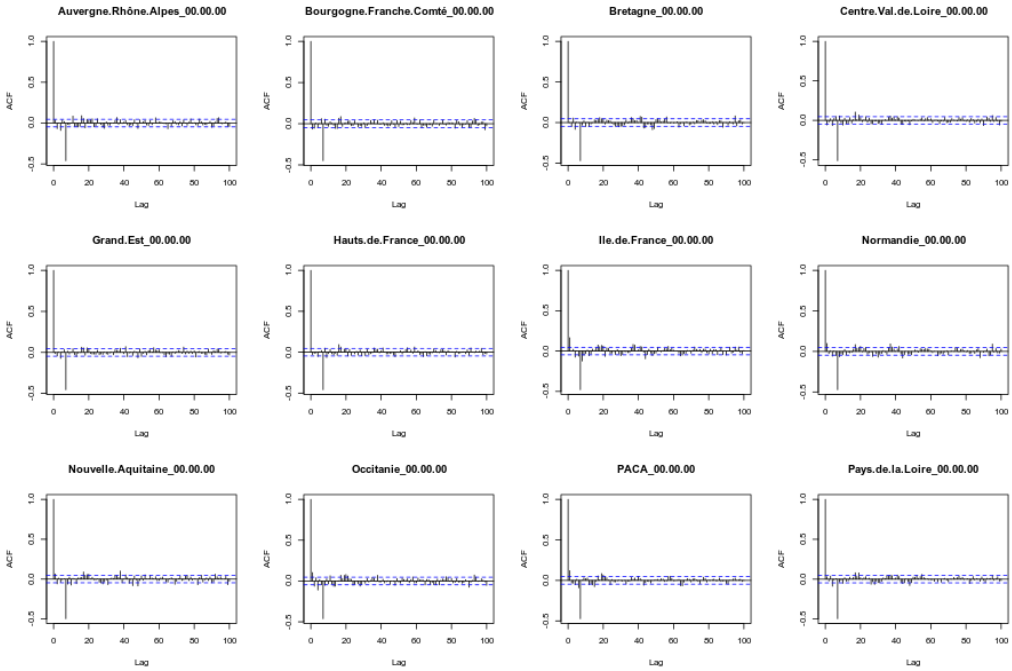

Figure 5: Autocorrelation function of the weekly differenciated series + another difference.

2.3 TODO GCC description

2.4 DONE GCC calculation

1. **DONE** Selecting k In order to select k, the maximum lag was taken by fitting autoregressive models to each of the series (using BIC). A maximum lag of 40 was used and was computed both in R and in Python. In both case, it found a maximum fitted lag of 37.

   - In R:

```
library(FitAR)

getOrder <- function(ts, order.max=40) {
  SelectModel(ts, ARModel = 'AR', Criterion = 'BIC', lag.max = order.max)[1,1]
}

k <- max(apply(consommation, 2, getOrder))
print(k)
```

```
[1] 37
```

   - In Python:

```
import statsmodels.api as sm
```

```
k = consommation.apply(
    lambda x: sm.tsa.arma_order_select_ic(
        x, ic='bic', trend='nc', max_ar=40, max_ma=1)['bic_min_order'][0]).max()
k
```

```
37
```

This lag seems appropriate when looking at the partial autocorrelation functions in figure 6, as that is where the last significant value is observed.

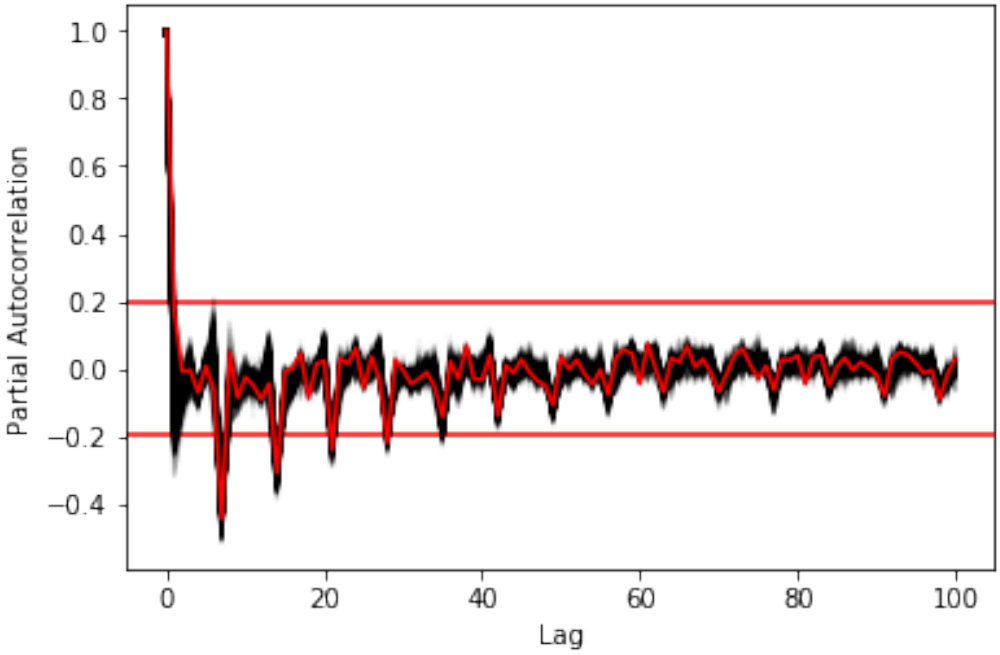

Figure 6: Partial autocorrelation of the stationary scaled data.

2. **DONE** Distance matrix The GCC was computed in both R and in Python to validate the results.

   - In R:

```
kMatrix <- function(ts, k) {
  m <- ts[1 : (length(ts) - k)]
  for (i in seq(k)) {
    m <- cbind(m, ts[(i+1) : (length(ts) - k + i)])
  }
  m
}

GCC <- function(ts1, ts2, k) {
  Xi <-  kMatrix(ts1, k)
```

```
  Xj <-  kMatrix(ts2, k)

  Xij <- cbind(Xi, Xj)

  det(cor(Xij))^(1/(k+1)) /
    (det(cor(Xi))^(1/(k+1)) * det(cor(Xj))^(1/(k+1)))
}
k<-37
combinations <- combn(dim(consommation)[2], 2)
DM_GCC <- matrix(0, dim(consommation)[2], dim(consommation)[2])
for (d in seq(dim(combinations)[2])) {
  distance <- GCC(consommation[, combinations[,d][1]],
                  consommation[, combinations[,d][2]], k)
  DM_GCC[combinations[,d][1], combinations[,d][2]] <- distance
  DM_GCC[combinations[,d][2], combinations[,d][1]] <- distance
}
rownames(DM_GCC) <- colnames(consommation)
colnames(DM_GCC) <- colnames(consommation)
write.csv(DM_GCC, file="data/DM_GCC_37_R.csv")
```

- In Python:

```
import numpy as np
from scipy.spatial.distance import pdist
from scipy.spatial.distance import squareform
import pickle


def k_matrix(ts, k):
    T = ts.shape[0]
    return np.array(
        [ts[(shift):T - k + shift] for shift in np.arange(0, k + 1)])


def get_GCC(ts1, ts2):
    k = 37
    Xi = k_matrix(ts1, k)
    Xj = k_matrix(ts2, k)
    Xij = np.concatenate((Xi, Xj))
    GCC = np.linalg.det(np.corrcoef(Xij)) ** (1 / (k + 1)) / (
        np.linalg.det(np.corrcoef(Xi)) ** (1 / (k + 1)) \
        * np.linalg.det(np.corrcoef(Xj)) ** (1 / (k + 1)) )
    return GCC


pdist_gcc = pdist(consommation.values.T, get_GCC)
DM_GCC = squareform(pdist_gcc)
DM_GCC = pd.DataFrame(
```

```
    DM_GCC, index=consommation.columns, columns=consommation.columns)
DM_GCC.to_csv('data/DM_GCC_37.csv')
```

The maximum difference between the results of the computation in the two language was of $\pm 5.3e^{-15}$ and can therefore be considered equivalent.

3 DONE Results[2/2]

3.1 DONE Clustering[2/2]

Hierarchical clustering was used, as it doesn't require a defined number of clusters to be set, and can directly be computed with a distance matrix.

3.1.1 DONE Linkage

More specifically, agglomerative clustering was used, where each data points starts in its own cluster and iteratively gets merged with its closest cluster. There are different methods to compute that intra-cluster distance, refered to as linkage method. The most popular methods were compared using the cophonetic correlation, which is the correlation coefficient between the distances between each point using their cluster distances and the original distance. A value closer to 1 means that the defined clusters respect better the original distances.

As such, both R and Python, the most conservative method was the average linkage and was therefore used to create the dendrogram (table 4). Different results were obtained for the 'centroid' and 'median' method, but still didn't beat the 0.77 of cophonetic correlation of the 'average' linkage.

Table 4: Cophonetic correlation of linkage methods.

	Average	Centroid	Complete	Median	Single	Ward	Weighted
Python	0.77	0.73	0.69	0.70	0.69	0.66	0.74
R	0.77	0.55	0.69	0.29	0.69	0.66	0.74

In 7 we can clearly see that there is a lot of structure. There are distances across the whole range of the GCC, making it easier to distinguish the groups. In fact, the regions appear the be the main influencing factor.

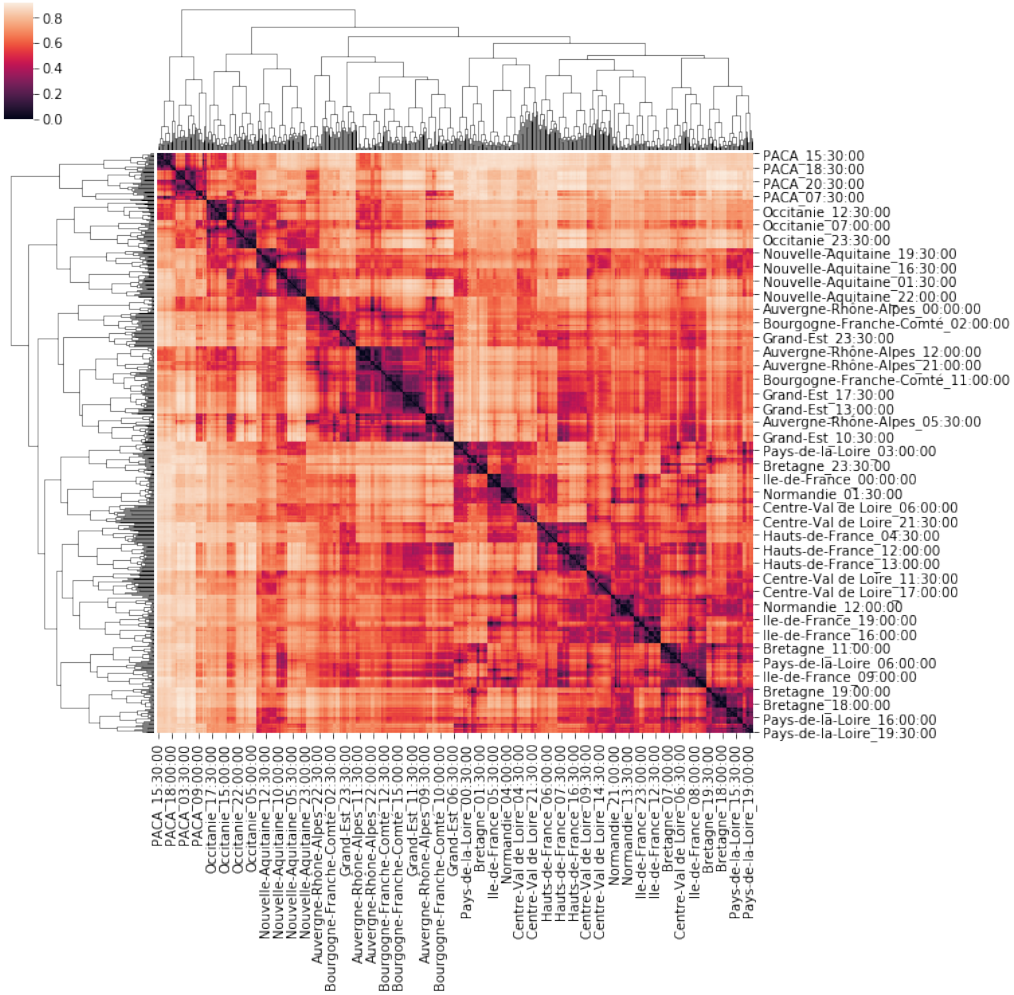

Figure 7: Heatmap of the distance matrix rearranged using the average linkage hierarchical clustering.

3.1.2 DONE Cluster number

Determining the number of cluster can be very challenging. The *factoextra* package in R provides functions to intent finding that number automatically. However, as you can see in figure 8, it isn't always that obvious.

The larger silhouette width is observed at 2 clusters but there is a small peak at 5 clusters. We can also see that the more clusters the better the gap statistic. However, we can see a small peak at k=5. Looking at the sum of square distance, we can also notice a small "elbow" at k=5.

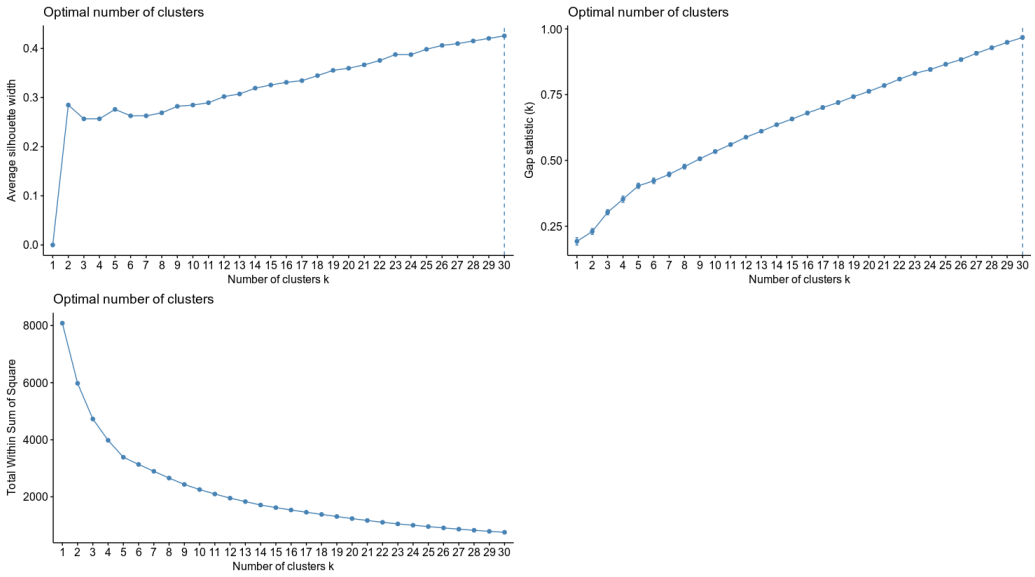

Figure 8: Mean silhouette width, gap statistic and total within cluster sum of square distance for each number of cluster.

This all suggest that there might be 5 clusters in our dataset, as shown on the dendrogram (fig. 9). Another way to look at those clusters is by looking the first 2 principal components of the distance matrix (fig. 10).

In fig. 11, we can see the silhouette width of each of the samples in their respective cluster. There seems to be some misclassification for some samples in cluster 3, but overall each cluster has significantly high silhouette width.

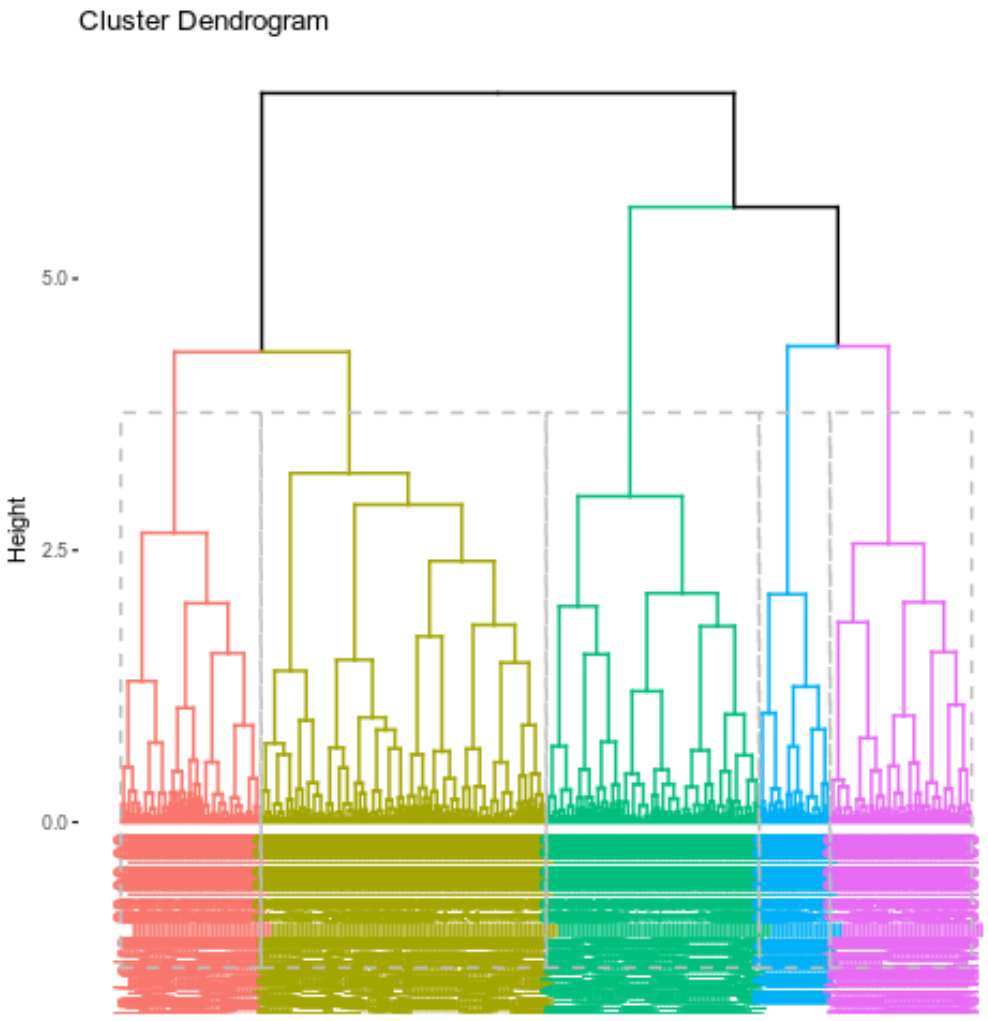

Figure 9: Dendrogram of the distance matrix using average linkage.

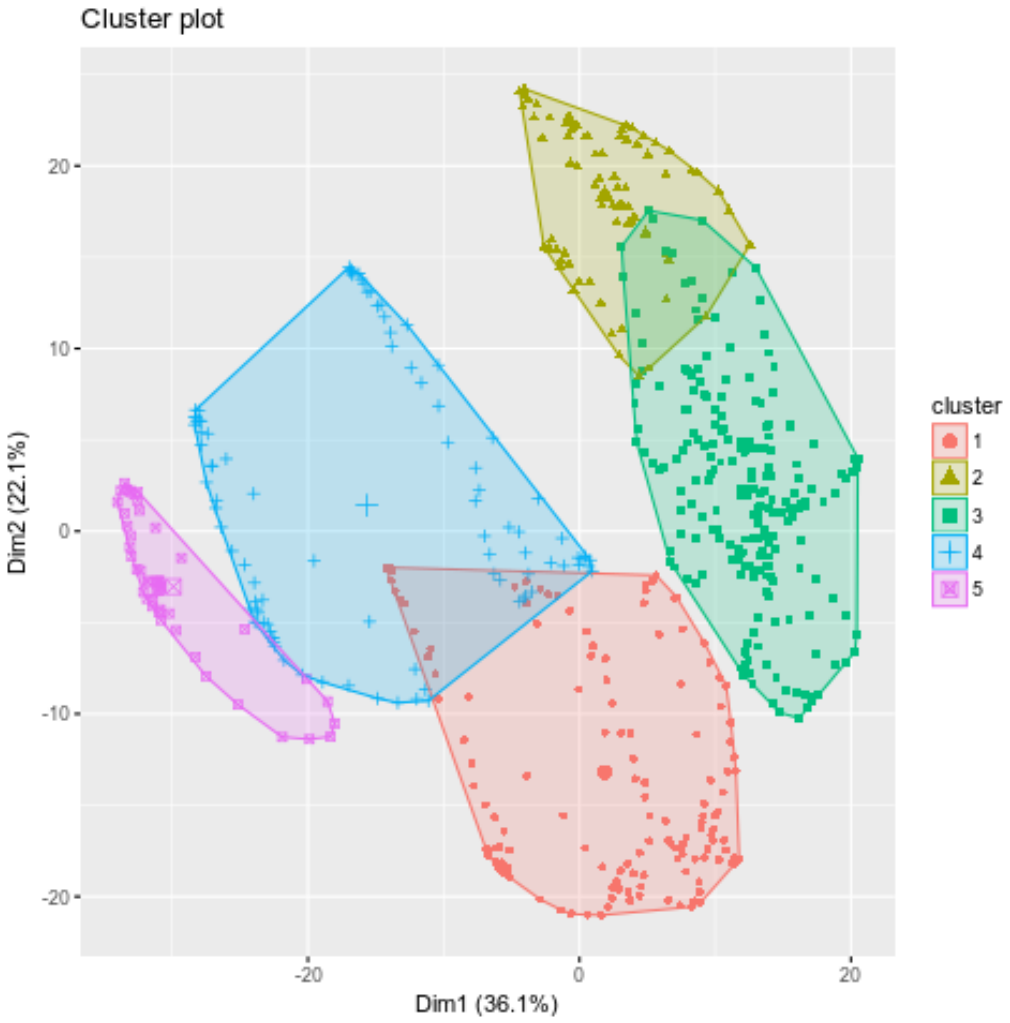

Figure 10: 5 clusters over the 2 principal components of the distance matrix.

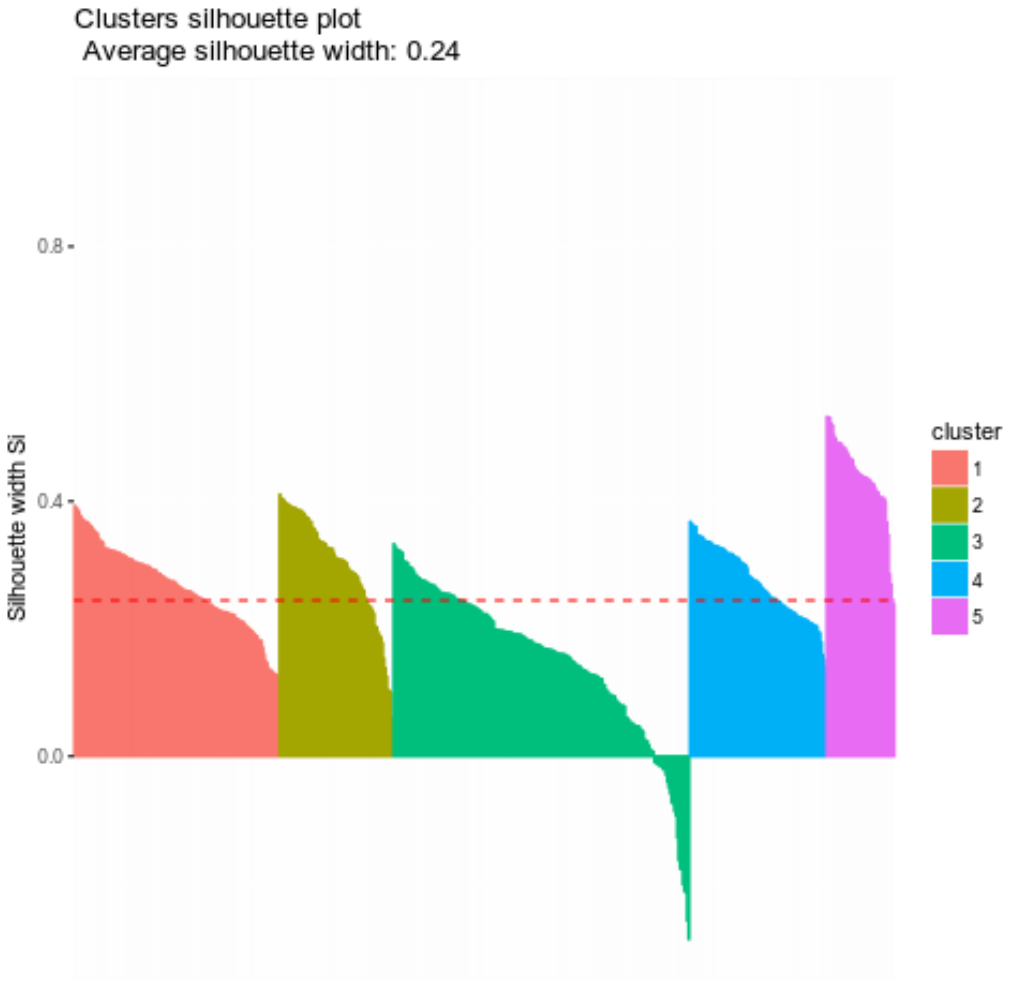

Figure 11: Silhouette width of the samples in each cluster.

3.2 DONE Cluster analysis[3/3]

3.2.1 DONE Mapping the clusters

If we were to only use 2 clusters, the PACA region is clearly the most distinct of all the regions (12).

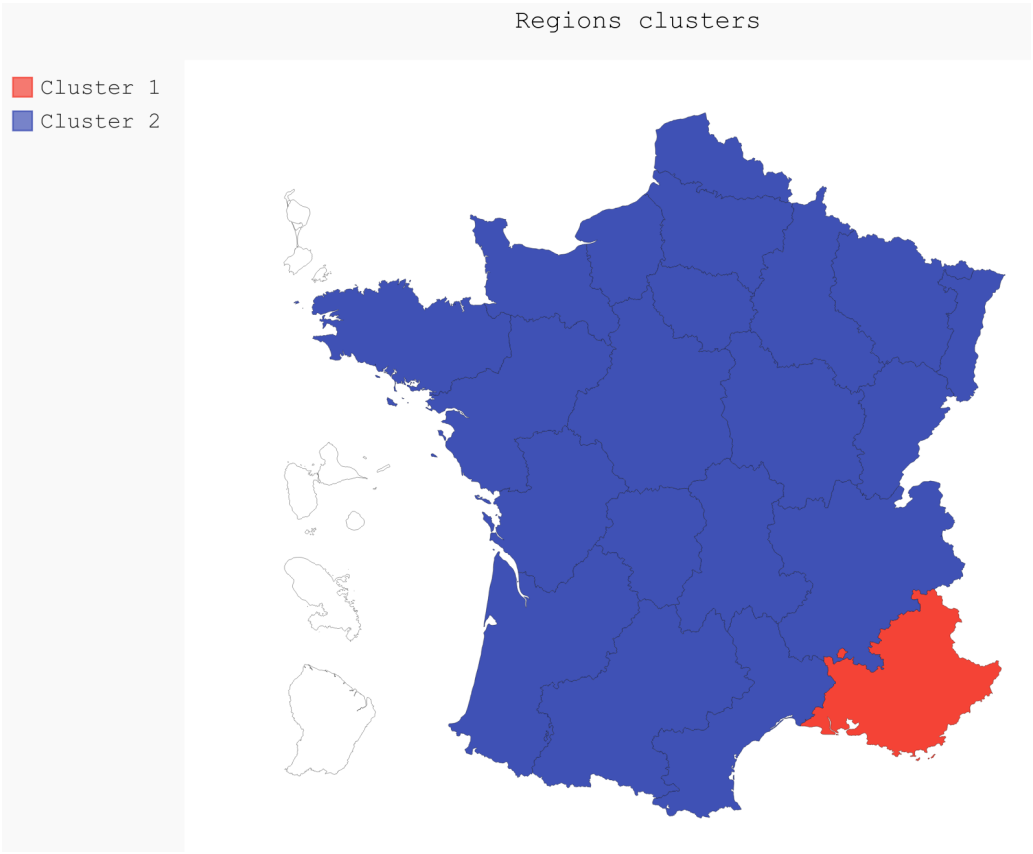

Figure 12: Map of the 2 clusters on the map of France. The regions shown are the old more numberous regions, but the boundaries of the 12 new reiongs are the same.

However, in order to have a deeper understanding of the composition of France, 5 clusters was the other clear delimitation. It is very clear here, that all the clusters have a strong geographical meaning. All regions are in different clusters apart from cluster 4 and 5 that are mixed geographically (table 5 and fig. 13), which are more defined by their consumption over time.

Table 5: Regions in each clusters.

1	2	3	4	5
PACA	N-A	A-R-A	Bretagne	Bretagne
	Occitanie	B-F-C	C-V-L	C-V-L
		G-E	I-F	I-F
			Normandie	Normandie
			P-L	P-L
				H-F

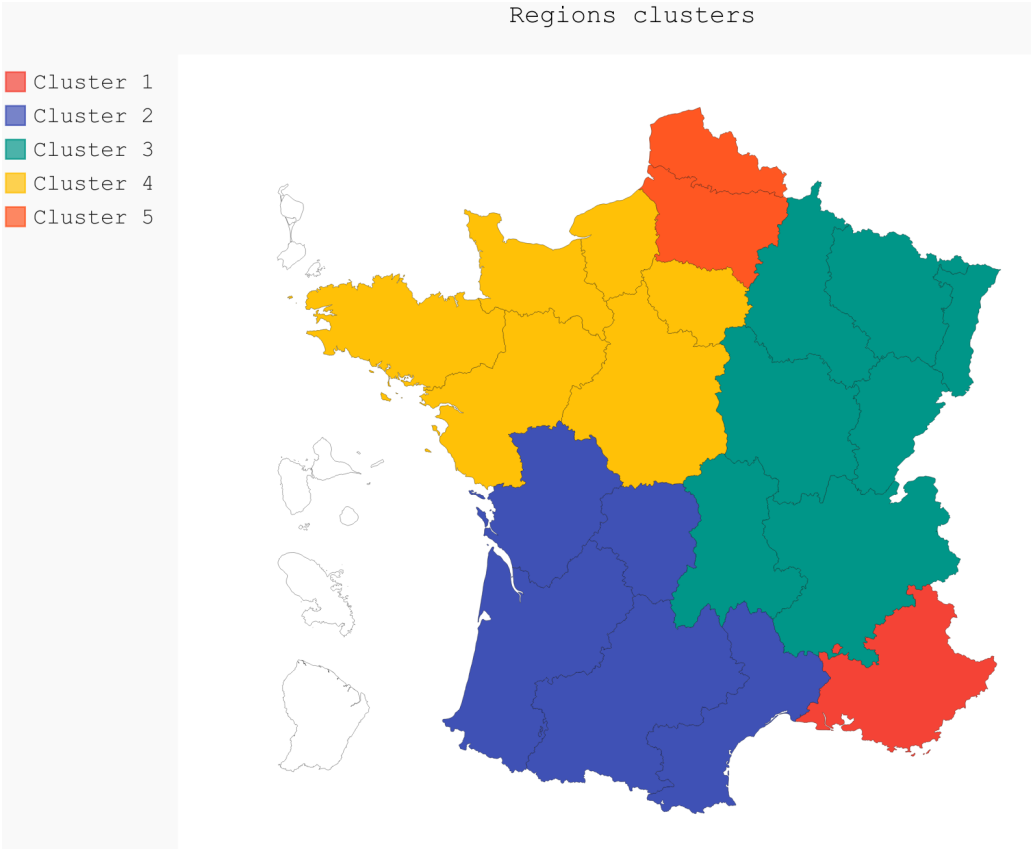

Figure 13: Map of the 5 clusters on the map of France. The regions shown are the old more numberous regions, but the boundaries of the 12 new reiongs are the same.

3.2.2 DONE Within clusters structure

In this section, the goal was to find out if there was more structure within each of the clusters. A dendrogram was plotted for each cluster and the label was coloured depending on the time of the day, where black is late in the day and red is early morning. The lighter colours are towards midday.

Cluster 1 only contains the PACA region. In figure 14, we can see that there are 3 main clusters, mornings from 6:30 to 11:00, midday-afternoon from 11:30 to 20:00, and the night cluster from 20:00 to 6:00. Days (11:30 to 20:00) and nights (20:30 to 11:00) are however the most well defined.

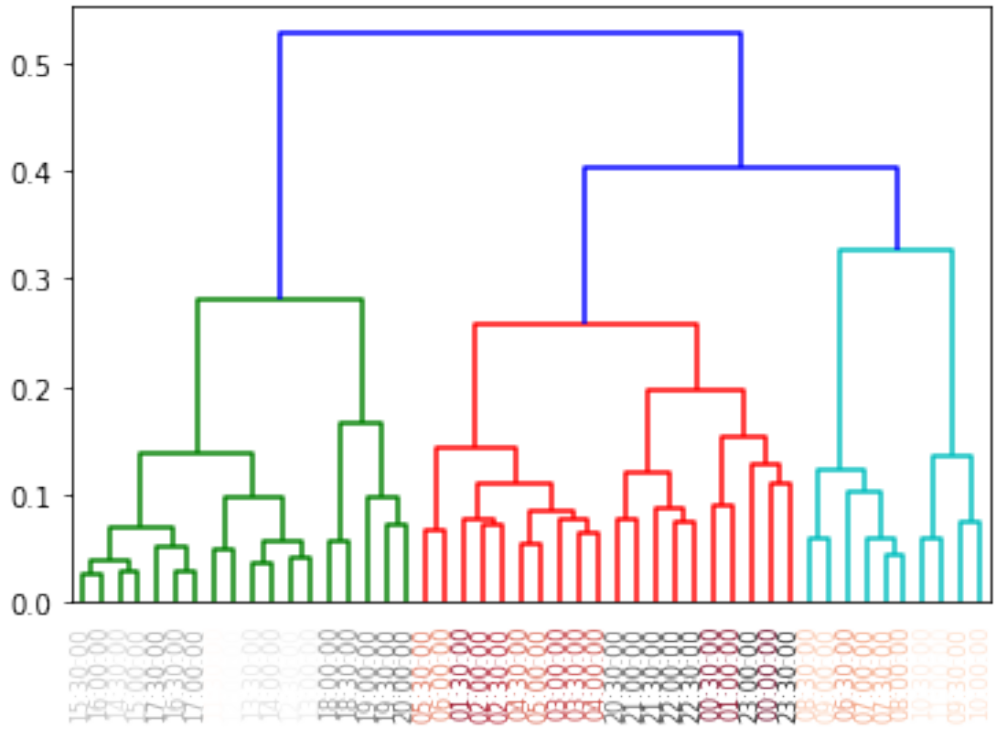

Figure 14: Dendrogram of cluster 1. Black is late in the day and red is early morning. The lighter colours are towards midday.

Cluster 2 contains 2 regions (Nouvelle-Aquitaine and Occitanie). In figure 15, in the top plot the label was coloured by the region and the bottom plot the label was coloured by the time of the day. We can see that the most important must important clustering is by region, but then similar clustering, by time of the day, as cluster 1 is observed.

In cluster 3, containing 3 regions (Auvergne-Rhône-Alpes, Bourgogne-Franche-Comté and Grand-Est) things are very different. The time of the day is the most important variable, as apart from Grand-Est, there are 2 main clusters, the late-night and early-morning cluster and the rest of the day (fig. 16).

Cluster 4 contains 4 regions (Bretagne, Centre-Val de Loire, Ile-de-France, Normandie and Pays-de-la-Loire), but only late night and early morning times . Here the regional clusters are very clear as all regions have been split with no clear time cluster (17).

In cluster 5, there are 5 regions, the same ones as in cluster 4 as well as Hauts-de-France.

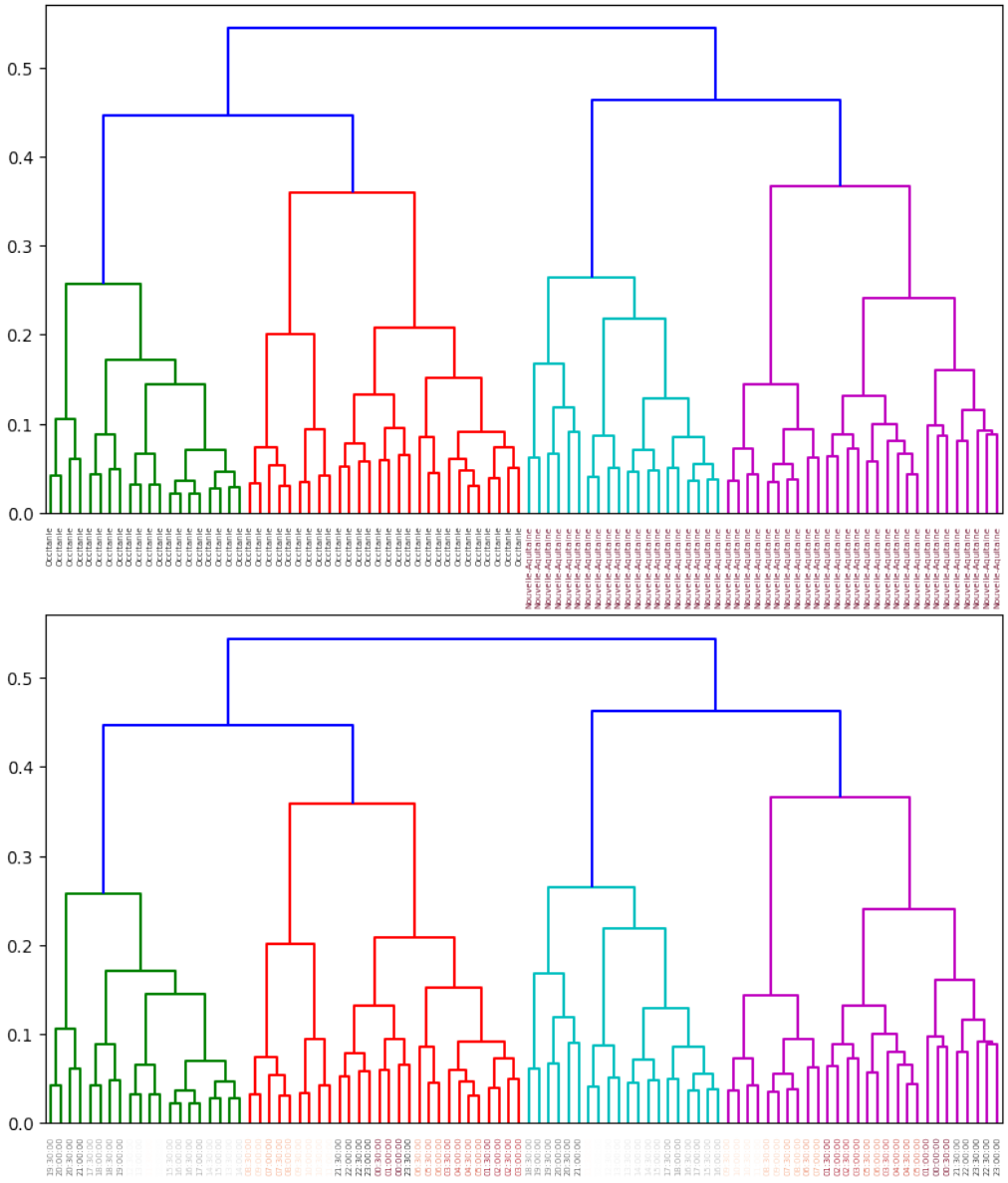

Figure 15: Dendrogram of cluster 2. Top: Black is Occitanie and red is Nouvelle-Aquitaine. Bottom: Black is late in the day and red is early morning. The lighter colours are towards midday.

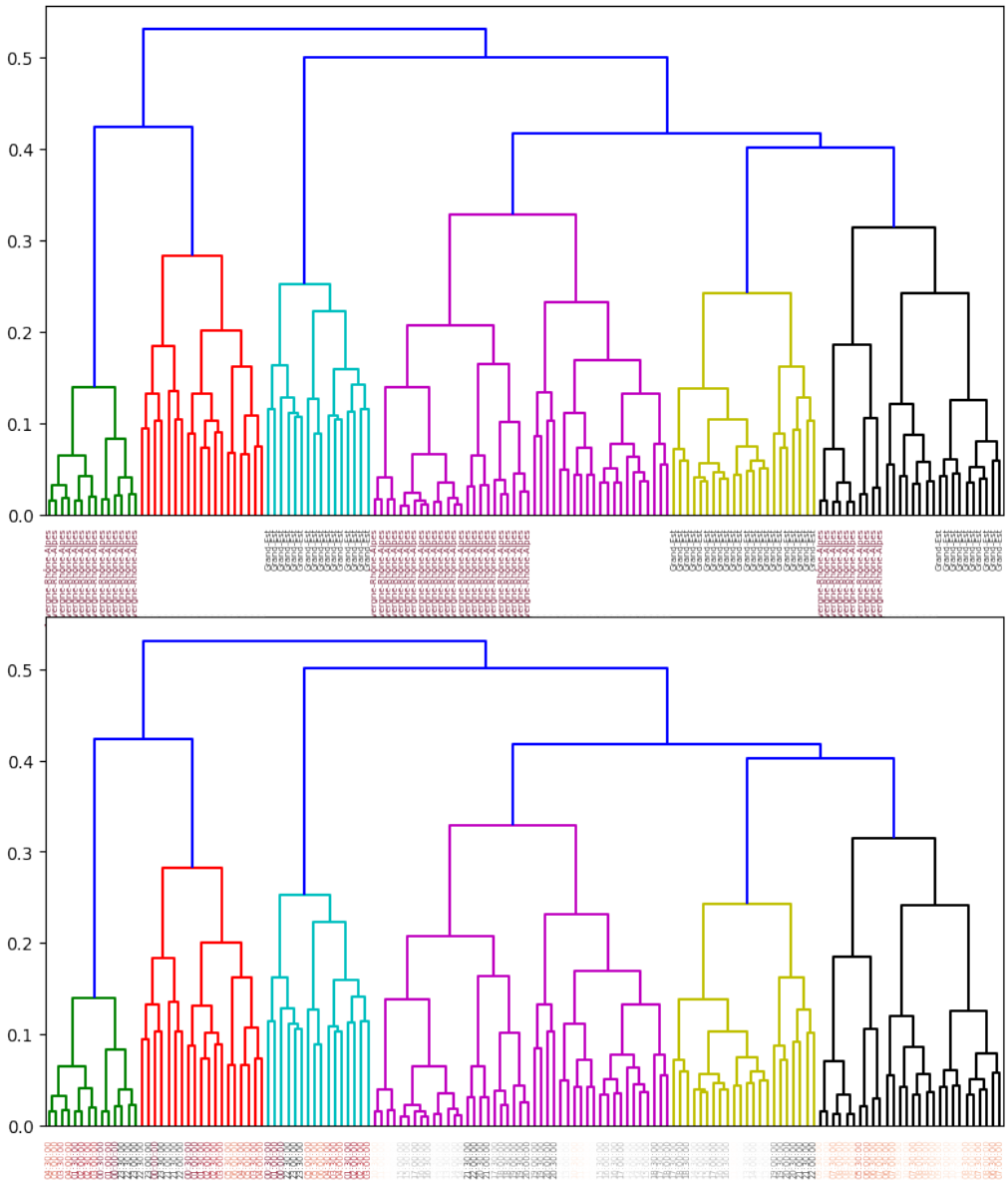

Figure 16: Dendrogram of cluster 3. Black is late in the day and red is early morning. The lighter colours are towards midday.

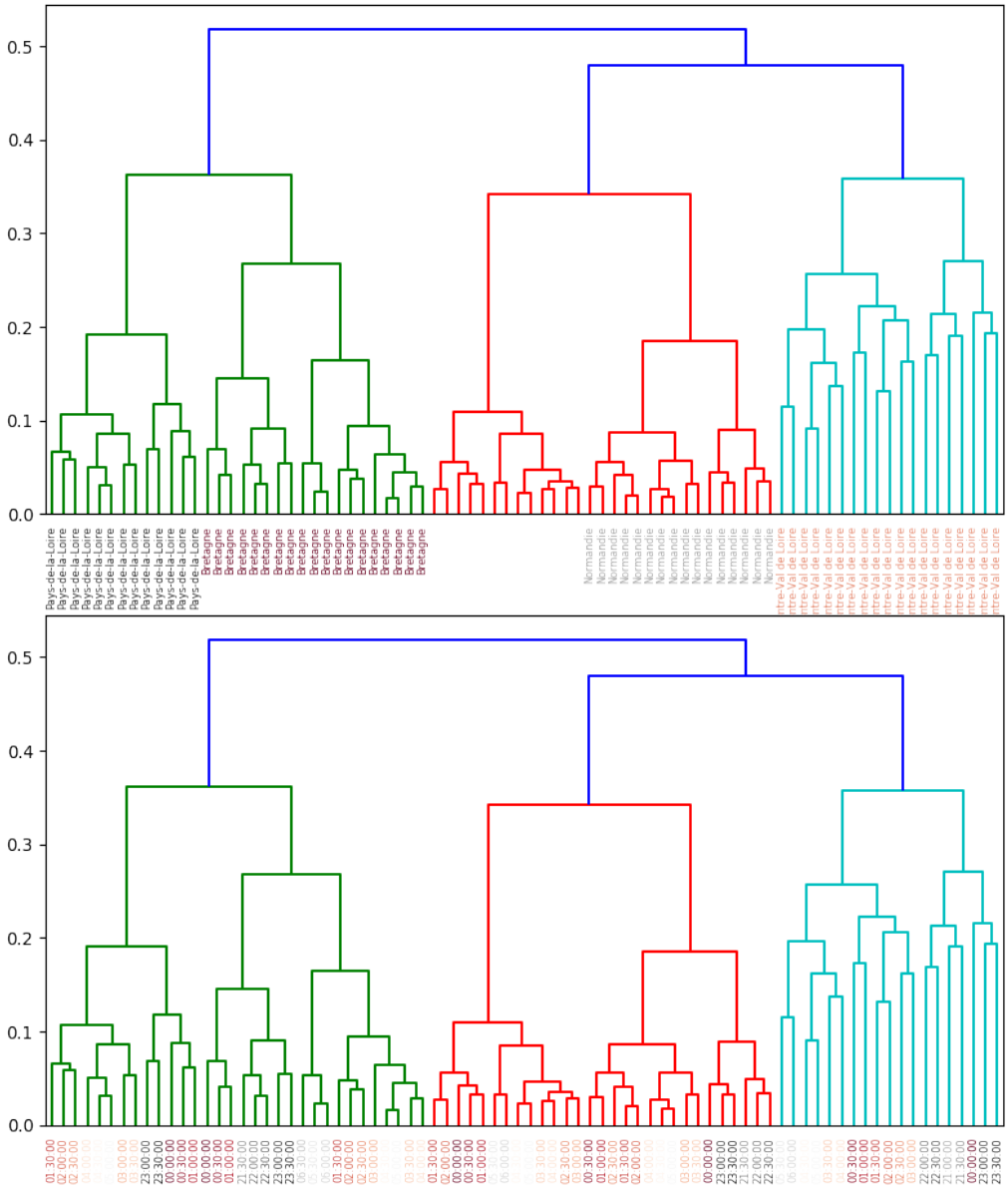

Figure 17: Dendrogram of cluster 4. Black is late in the day and red is early morning. The lighter colours are towards midday.

At all times, the Hauts-de-France was grouped with the evenings of the Centre-Val-de-Loire, Normandie and Iles-de-France, whereas the other regions are clustered into mornings and evenings.

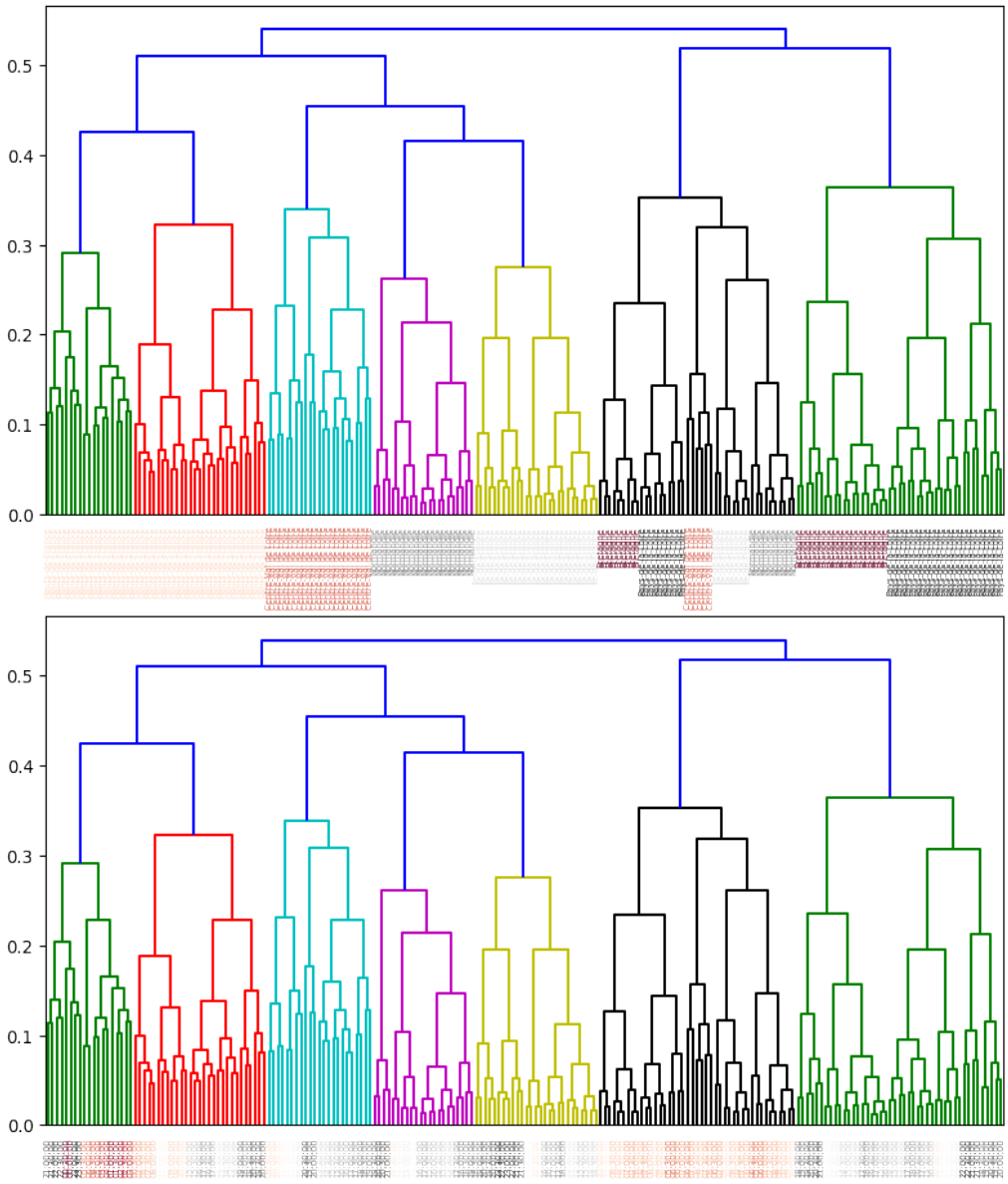

Figure 18: Dendrogram of cluster 5. Black is late in the day and red is early morning. The lighter colours are towards midday.

3.2.3 DONE Clusters trends

As no information about the size of the population in each region was used, the absolute consumption were not compared between clusters. However, we can still compare relative changes over the years (fig. 19), seasons (fig. 20) and a typical day (fig. 21).

The 1 year trends of each cluster seem to suggest that the regions that had lower consumptions in 2013-2014 have increased their consumptions in 2016-2017, and inversely for regions that had it higher in the 2013-2014 period (fig. 19). The PACA region (cluster 1) is also clearly differenciated from the other ones. However, it is difficult to get clear conclusions as there are not enough data to analyse long term trends.

Figure 19: 1 year moving average trend of each cluster.

In the 3 months trend (fig. 20), we can see that cluster 1 and 2 have a higher energy consumption during the summer. This is most likely due to the use of air conditioning, as those 2 clusters are in the south of France, which is not really common (nor necessary) in the north.

Over the day (fig. 21), cluster 1, and to a smaller extend cluster 2, tend to use electricity later than the other regions. Again, this is most likely due to the different life style between the north and south regions of France. As it is very warm during the days, people tend to go out more in the evenings, as shown by the higher consumption around 20:00.

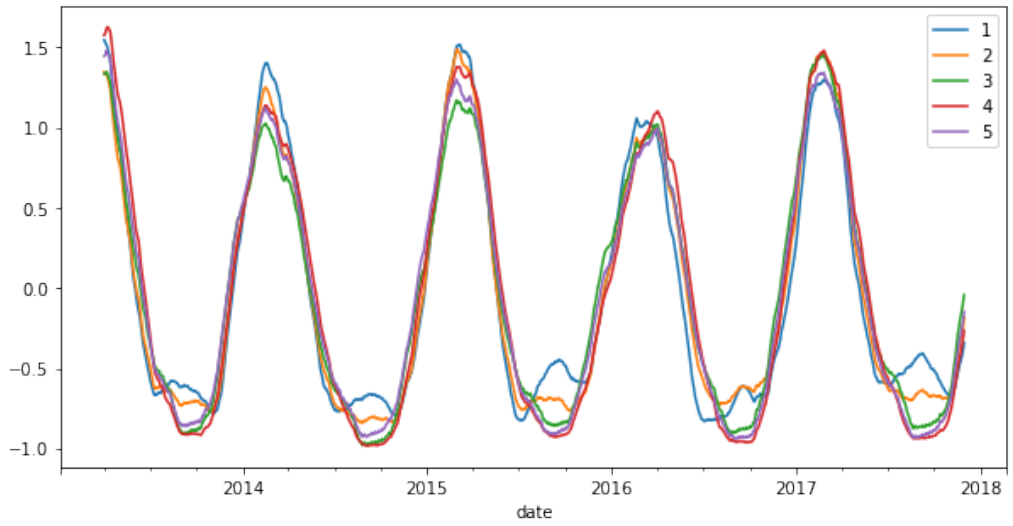

Figure 20: 3 months moving average trend of each cluster.

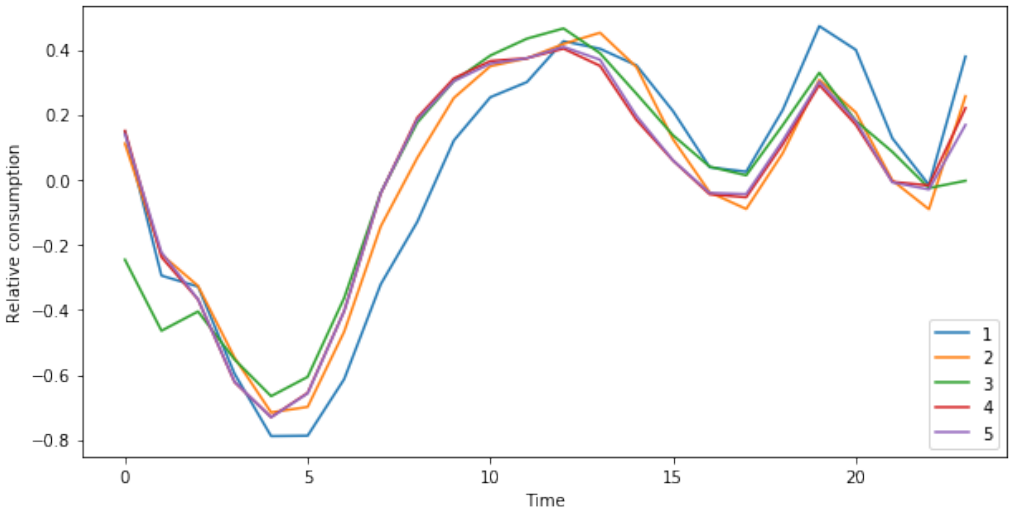

Figure 21: Hourly mean consumption of everyday for each cluster.

4 TODO Conclusion

References

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C. (1988). Algorithms for clustering data.