France regional electricity consumption clustering using Generalised Cross Correlation.

Pierre Mercatoris

<2018-05-19 Sat>

1 Introduction

1.1 Cluster electricity consumption using GCC

Jain and Dubes (1988)

1.2 Clustering time series

2 Methodology

2.1 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 Data preparation

2.2.1 Cleaning

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

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

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

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
.....			

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 there are mostly only data every 30 minutes, the table was resampled by taking the sum for each 30 minutes, resulting in the table below (table 2).

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	

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).

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.

2.2.2 Transformation

1. Stationarity

   The original series feature a strong seasonality as show in figure 3.

   To try and remove it, I have taken the weekly difference (difference between all the values separated by 7 days). Now there is still some correlation, but it is better (fig. 4).

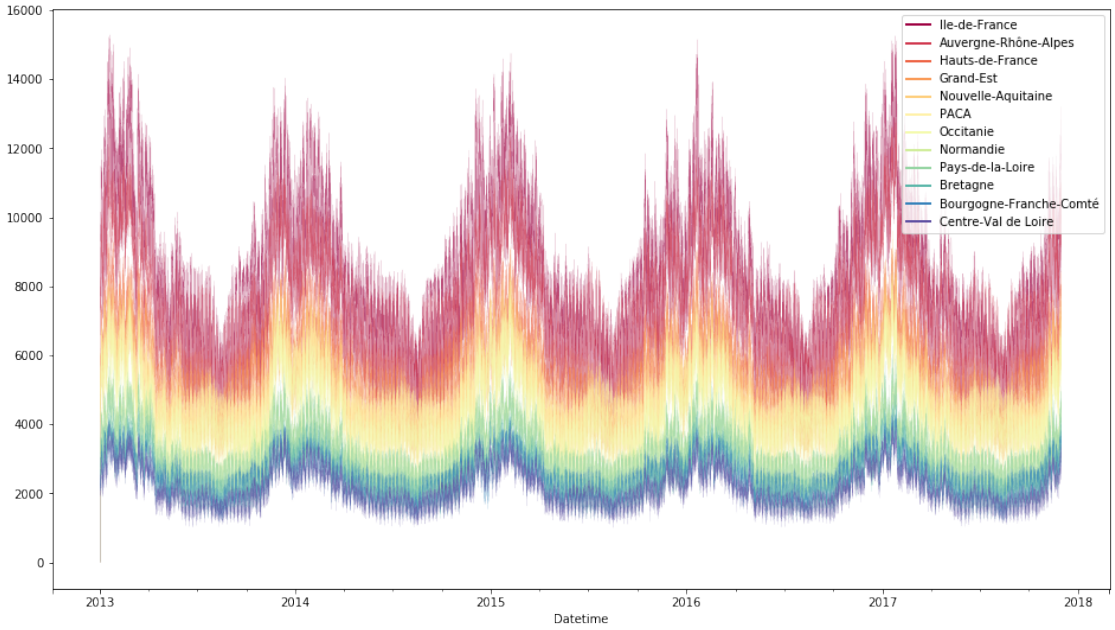


Figure 1: Mean electricity consumption of each of the french regions from 2013 to end 2017.

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

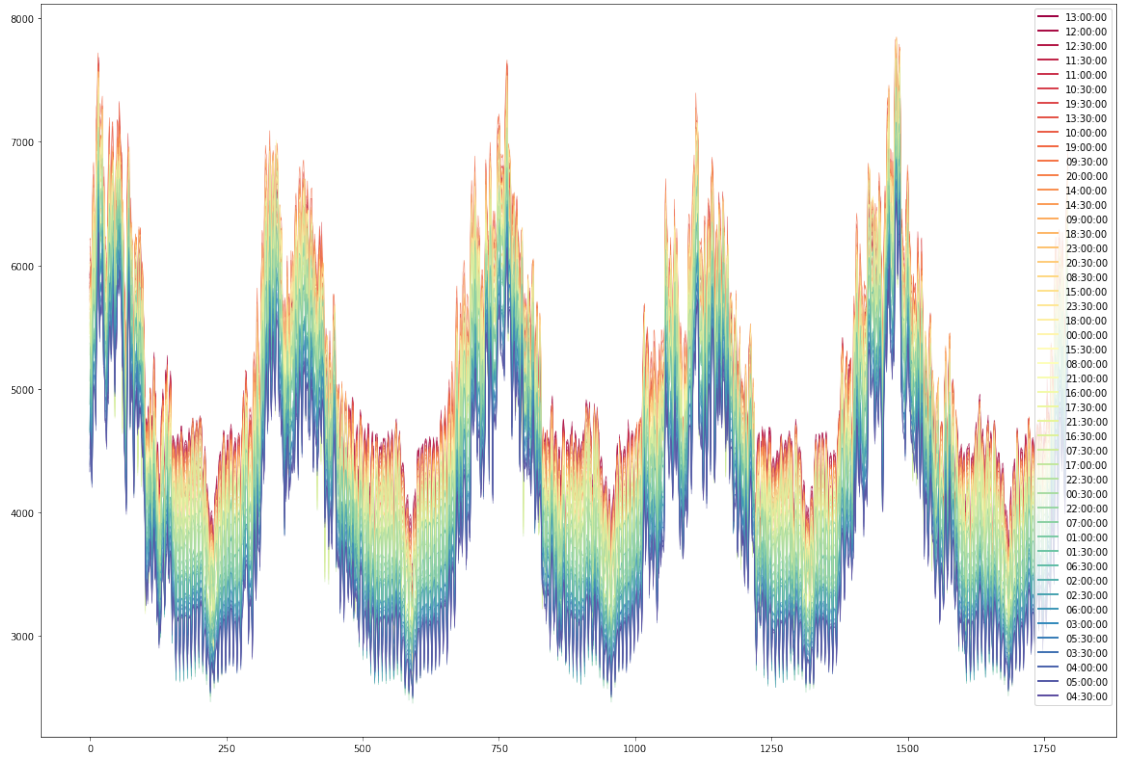


Figure 2: Mean electricity consumption for all the regions of France at different times.

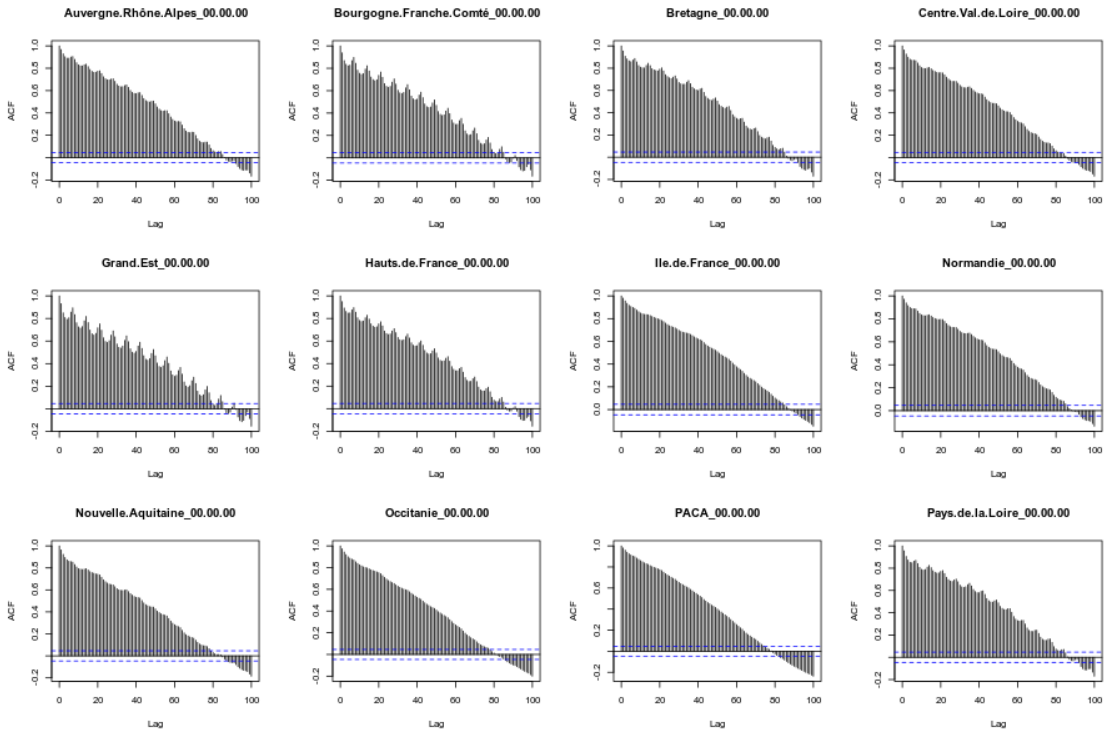


Figure 3: Autocorrelation function of the original data.

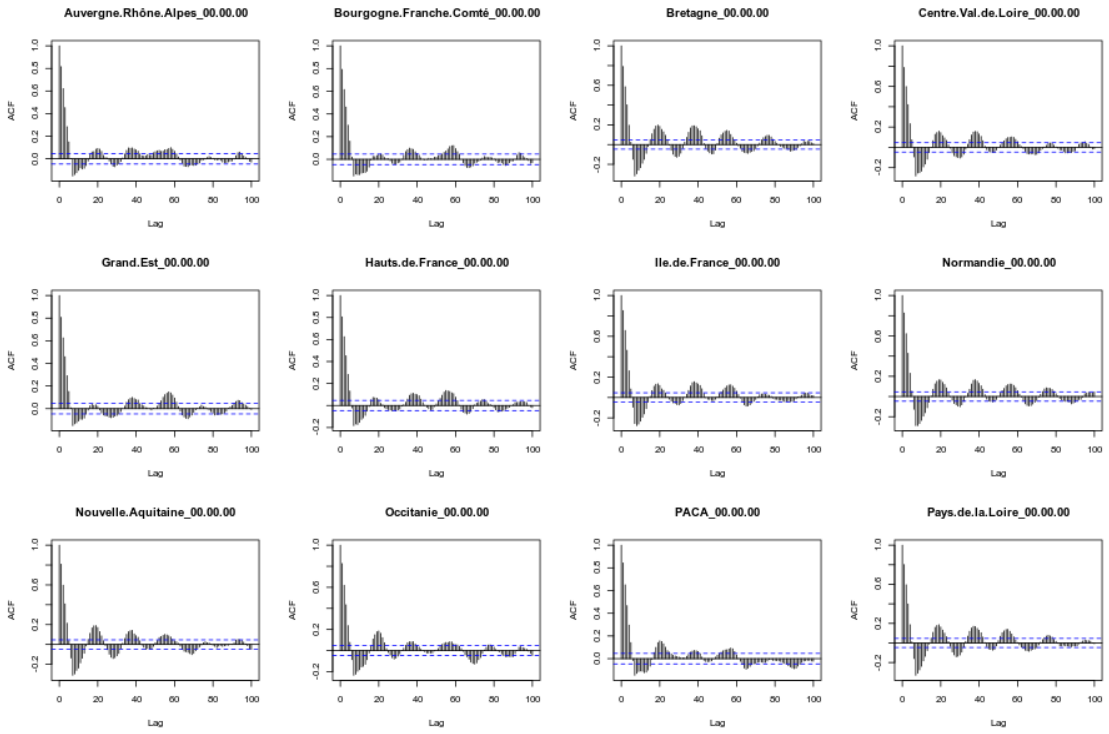


Figure 4: Autocorrelation function of the weekly differenciated series.

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

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

2. Standardisation

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

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

2.3 GCC description

2.4 Distance calculation

1. Selecting k
2. Distance matrix

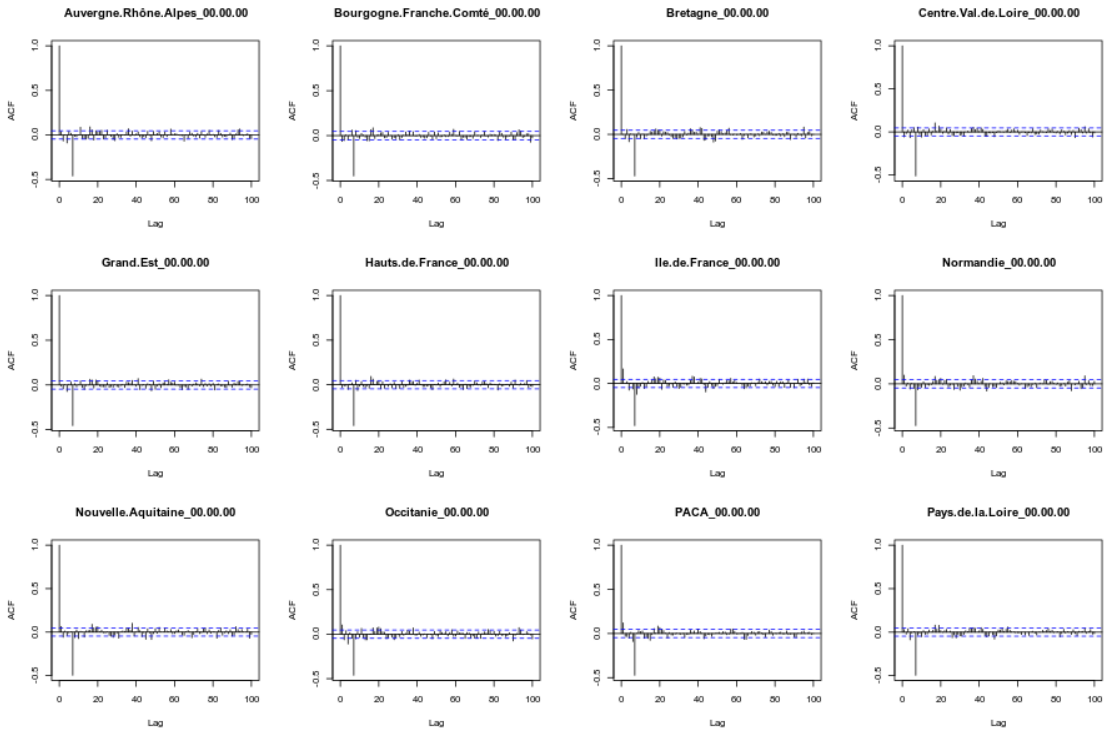


Figure 5: Autocorrelation function of the weekly differenciated series.

3 Results

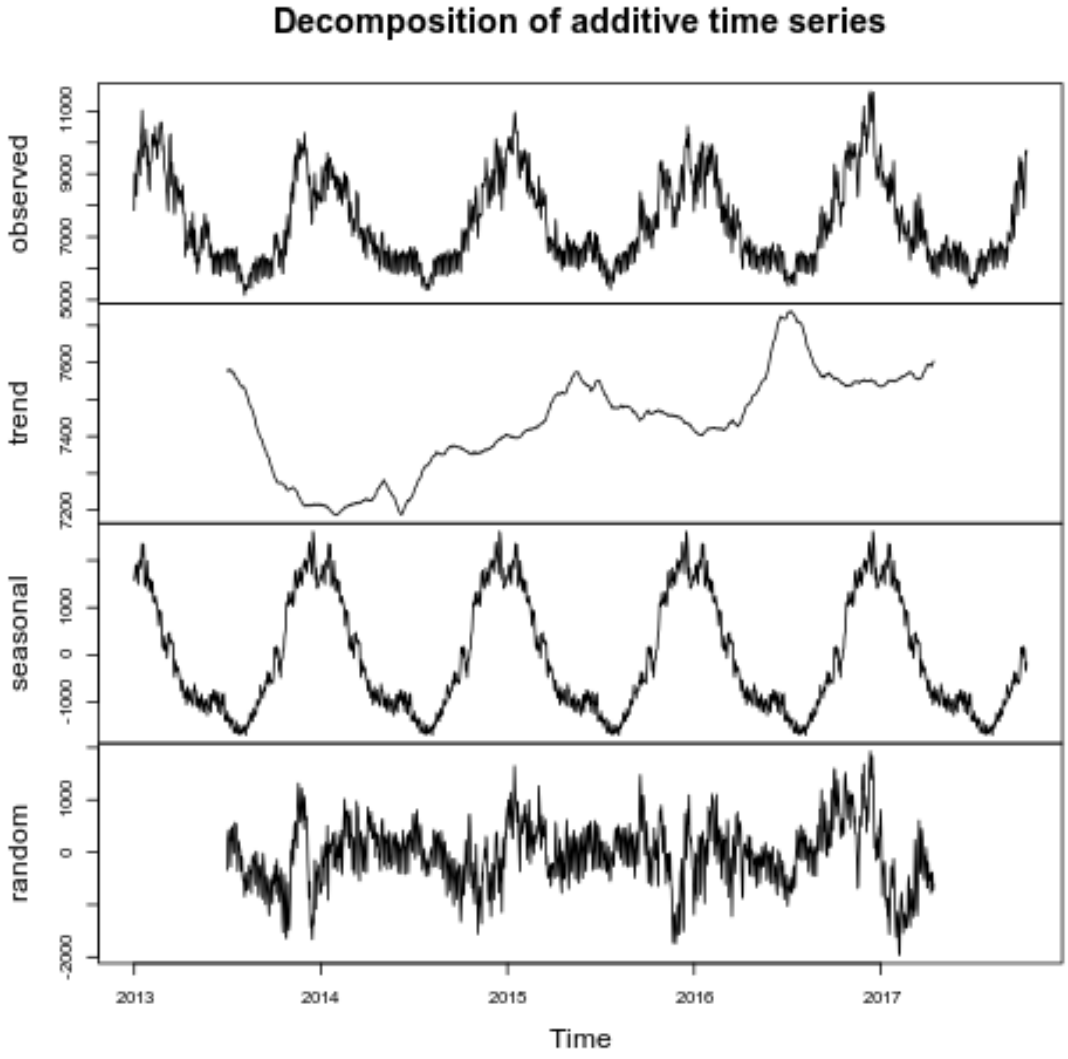
3.1 Clustering

3.2 Cluster analysis

```
library(tidyverse)
library(xts)

consommation <- read.csv('./data/consommation.csv', row.names='date')

ts1 = ts(consommation[,1], frequency = 375, start = 2013)
plot(decompose(ts1))
```



4 Conclusion

References

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