Ecole Nationale des Ponts et Chaussées


# Hydrogen IMI Project

Baptiste Vert

May 12, 2024

# Contents

# 1 Introduction : presentation of the project

## 1.1 Motivation of the project

The situation is the following: we have an electrolyser which produce hydrogen from electricity. We have two sources of electricity; one is a renewable energy park composed of wind turbines and solar panels and the other is the grid. The first one can symbolise free power, that means the power infrastructures are part of the infrastructure of the company, it can also symbolise costly power but at a negotiated price through a PPA. In both cases it symbolises the prioritized source of electricity, but with a lot of uncertainty coming from the renewable part of the power. The other source of electricity namely the grid would then represent some infinite source of electricity whose price would change over time. The idea here is to use the storage of hydrogen to balance the uncertainty of the renewable energy park, indeed storing cheap or clean electricity in batteries is not efficient, since there is dissipation over time, but the storage of hydrogen does not present this drawback. So, when a lot of renewable electricity is available, it is stored as hydrogen.

The project is based on these considerations. There is an hourly demand for hydrogen, an electrolyser, a tank to store the hydrogen, a battery, a renewable electricity park and a grid. To fulfil the demand the electrolyser needs electricity to produce hydrogen, it uses in priority the one from the park, then if it needs more than what the park can provide, it completes with the grid, when the demand is fulfilled, there are different options. First, let us imagine that the park was able to provide a lot of electricity during the first hour, but during the following hour it can only provide little electricity so not enough to fulfil the demand of this hour by using only electricity from the park, so it has to use electricity from the grid which is at this hour very expensive. Therefore, a better production plan would have been to produce more hydrogen during the first hour and to store the surplus in the tank.

Moreover, there are some capacity constraints. The tank cannot contain more hydrogen than some fixed limit, the electrolyser cannot produce more hydrogen per hour than some other fixed limit, hence the interest to have a battery. Indeed, let us imagine the quantity of electricity from the renewable park available during the first hour was gigantic, so the electrolyser produced as much hydrogen as it was able to, a part was used to fulfil the demand and the surplus was stored in the tank. But to still benefit from this gigantic quantity of free/very cheap electricity, the battery is used to store as much electricity as it can, although it would have been more efficient to store this energy as hydrogen if it had been possible.

Now, the most important problem of this model is the uncertainty, indeed it is always possible to make predictions about the energy available at the following hour or even at the following days, this is what is behind the elaboration of the day-ahead price for electricity. But when it comes to weeks or even months, it gets harder and harder to have correct predictions, so if the production plan was made for different consecutive weeks, the costs of production, which here correspond to the prices at which electricity was bought from the grid, since in the following modelling we will not consider any other OPEX and CAPEX, could be much higher than in the optimal case where the predictions were true.

Consequently, a possible solution, which will be studied in this project, would be to constraint some final levels of stored hydrogen. To be more precise let us imagine that a production plan is made over four consecutive weeks, to make this plan a certain prediction model for the renewable energy available at each hour is used, but since they are only predictions, and the idea is to balance the uncertainty, the levels of hydrogen reached after the last hour of the first three weeks are constrained by a lower bound. This idea is linked with the concept that production plans are not made for being resilient over time, so when the different variables are settled like the hourly working of the electrolyser, it cannot be

changed, hence the idea to have an optimal production plan that can reduce the impact on the costs of the uncertainty associated with the renewable energies. But to keep a feasible production plan we will allow to change the source of energy at each hour, that means if we predicted more electricity from the renewable park that was actually available, we can complete with electricity from the grid, since it is considered as an infinite source of power.

## 1.2 Construction of the optimal production plan

The inputs here are the period over which we want to build our production plan, which is defined by the values of some hourly variables, such as the quantity of electricity used from the renewable park, the quantity of electricity used by the electrolyser, etc. The period could be 4 weeks, the other inputs are the hourly demand, the different capacities: electrolyser, tank, battery, solar capacity wind capacity, the predicted hourly prices of energy for the period and the predicted quantities of wind and solar power available at each hour (obtained that predicting the profiles of these energy and then multiplying it by their capacities). Since the function to minimize is a sum of fixed cost and variables (here the quantity of energy from the grid), and since the constraints are all linear (for example fulfilling the demand), this problem is nothing else than a linear optimization problem, which can be solved very efficiently, i.e., finding the optimal solution very quickly. Furthermore, the infrastructure was designed such that it could always fulfil the demand, that is the different capacities are large enough to make sure the hourly quantity of available hydrogen (from the tank or the electrolyser) is always greater than the hourly demand.

Nonetheless, there can be different variants of this linear optimization problem, for example cost for hourly production change could be added to have a smoother working of the infrastructure, the energy from the park could be either considered as free or with a low cost but that could be at some hours higher than the one of the grid, but then it would have to been ensured that it is still the electricity from the park which is used, for example by adding curtailing cost for the unused energy from the park.

## 1.3 Presentation of the study of the impact of the final levels of hydrogen

The optimal production plan with the final levels of hydrogen at the end of each week and the prices and energy profile is now made. It is therefore instantly applied for the real weeks, but as seen previously predictions can turn false, for example the predicted prices are not the ones on the market, so the predicted optimal cost would be different from the real one, but it could go in the two directions. For instance, suddenly a lot more of renewable energy than we predicted is available, so there is no need to use electricity form the grid, which results in a better cost. So, to compute a real cost, we need to transpose our production plan into the real case, but as explained before in this modelling we consider that the only part of the production plan that can be changed is the origin of the electricity used by the battery and the electrolyser, but since we also consider that the grid is an infinite source of energy we can just look at the hourly aggregated needed quantity of electricity (from the renewable park and the grid), then use as much electricity from the renewable park as possible and then finally complete with the grid, which will give the real cost of this production plan in the real case.

Now, the issue is the different values of the final levels of hydrogen. A first approach to get an idea is to generate random levels and see their impact on the final cost in the real case. Then some heuristics could be implemented to get a better approach of the optimal levels of stock, and finally a statistics

study of these different levels could be study via R, through the estimation of a regression function.

### 1.4 Presentation of the study of the data used in the problem

This study was made using R studio, it relies on times series analysis theory.

The database from ENTSO-E Transparency is the one used for this project, it contains data related to renewable energy (wind and solar) for different European countries. A part of the study was focused on Germany, the solar energy generation time series presents significant structures with some pattern and seasonality. It is less the case for the wind energy and for the prices. For the price it is probably linked to the fact that it is built with the offer of energy but also with the demand, which may not always be strongly correlated.

## 2 Time series analysis for renewable energies

### 2.1 Presentation of the data

License and attribution of the data : "Open Power System Data. 2020. Data Package Time series. Version 2020-10-06. https://doi.org/10.25832/time_series/2020-10-06. (Primary data from various sources, for a complete list see URL)."

Source : ENTSO-E Transparency

"This data package contains different kinds of timeseries data relevant for power system modelling, namely electricity prices, electricity consumption (load) as well as wind and solar power generation and capacities. The data is aggregated either by country, control area or bidding zone. Geographical coverage includes the EU and some neighbouring countries. All variables are provided in hourly resolution. This package version only contains data provided by TSOs and power exchanges via ENTSO-E Transparency, covering the period 2015-mid 2020."

"The data package covers the geographical region of 32 European countries."

Example of Germany :

- **DE_load_actual_entsoe_transparency**
  - **Type**: Number
  - **Description**: Total load in Germany in MW as published on ENTSO-E Transparency Platform
- **DE_load_forecast_entsoe_transparency**
  - **Type**: Number
  - **Description**: Day-ahead load forecast in Germany in MW as published on ENTSO-E Transparency Platform
- **DE_solar_capacity**
  - **Type**: Number
  - **Description**: Electrical capacity of solar in Germany in MW

- **DE_solar_generation_actual**
  - **Type**: Number
  - **Description**: Actual solar generation in Germany in MW
- **DE_solar_profile**
  - **Type**: Number
  - **Description**: Share of solar capacity producing in Germany
- **DE_wind_capacity**
  - **Type**: Number
  - **Description**: Electrical capacity of wind in Germany in MW
- **DE_wind_generation_actual**
  - **Type**: Number
  - **Description**: Actual wind generation in Germany in MW
- **DE_wind_profile**
  - **Type**: Number
  - **Description**: Share of wind capacity producing in Germany
- **DE_wind_offshore_capacity**
  - **Type**: Number
  - **Description**: Electrical capacity of wind_offshore in Germany in MW
- **DE_wind_offshore_generation_actual**
  - **Type**: Number
  - **Description**: Actual wind_offshore generation in Germany in MW
- **DE_wind_offshore_profile**
  - **Type**: Number
  - **Description**: Share of wind_offshore capacity producing in Germany
- **DE_wind_onshore_capacity**
  - **Type**: Number
  - **Description**: Electrical capacity of wind_onshore in Germany in MW
- **DE_wind_onshore_generation_actual**
  - **Type**: Number
  - **Description**: Actual wind_onshore generation in Germany in MW
- **DE_wind_onshore_profile**

- **Type**: Number
- **Description**: Share of wind_onshore capacity producing in Germany

So we can see that for i,stance the Germany column misses a price_day_ahead line, this is due to the bidding zone.

A bidding zone is the largest geographical area within which market participants can exchange energy without the need to attribute cross-zonal capacity. Currently, bidding zones in Europe are primarily defined by national borders. These zones facilitate electricity trading within a specific region.

Why the Need to Review Bidding Zones?

The existing European electricity target model requires defining bidding zones based on network congestions. The goal is to maximize economic efficiency and cross-zonal trading opportunities while ensuring security of supply. By reviewing bidding zones, we can identify structural congestions and analyze different bidding zone configurations.

Benefits of Well-Defined Bidding Zones:

- Increased Opportunities for Cross-Zonal Trade: Clearer bidding zone configurations enhance cross-border electricity transactions.
- Efficient Network Investments: Properly defined zones allow for better planning and investment in transmission infrastructure.
- Cost-Efficient Integration of New Technologies: Improved bidding zones facilitate the integration of renewable energy sources and other innovations.

Therefore the database only contains the price associated to bidding zone, and Germany has over the past year been in two different bidding zone, the first one was composed of Germany, Luxembourg and Austria, and the second one was composed of only Germany and Luxembourg. So it is the column DE_LU_price_day_ahead which contains the price of the electricity for Germany after 30/09/2017.

Consequently, and since the number of ways of production of renewable energies has increased these past few years, we decided to restrict from 30/09/2017 to 30/09/2020.

Here is an explanation concerning the line price_day_ahead. We can also add, that the negative prices observed in the database make sense, indeed let us imagine the electricity producers know that tomorrow they will produce a lot of megawatts because of great weather conditions, so instead of stopping the production because they would produce more energy that it would be needed, they prefer to pay to have their production bought, hence the negative price.

So the **Day-Ahead spot price** for electricity is given for the **next day** (instead of the current day) for several important reasons:

1. **Production and Consumption Planning**:

- Electricity producers and grid operators need to plan electricity production based on the expected demand.
- Knowing the price in advance helps producers decide which power plants to turn on or off, based on profitability.

2. **Electricity Market**:
   - The electricity market operates as a wholesale market where producers sell electricity through auctions.
   - Buyers (electricity suppliers, businesses, etc.) want to know prices in advance to make informed decisions.
3. **Network Balancing**:
   - Grid operators must balance real-time electricity supply and demand.
   - Knowing prices in advance helps anticipate demand fluctuations and adjust production accordingly.
4. **Cost Optimization**:
   - Companies and industrial consumers can plan activities based on anticipated prices.
   - They can reduce costs by using electricity when prices are lower.

In summary, the Day-Ahead spot price is essential for grid stability, planning, and economic efficiency.

The solar capacity of an electrical grid refers to the total amount of electrical energy that solar photovoltaic (PV) installations can generate and inject into the grid. Specifically:

- It represents the maximum power that solar panels can produce under optimal conditions (e.g., full sunlight).
- It is expressed in kilowatts (kW) or megawatts (MW).
- Solar capacity is a key indicator for assessing the contribution of solar energy to the energy mix of a grid.

In summary, solar capacity represents the installed power of solar panels connected to the electrical grid, ready to produce electricity when the sun shines.

Actually concerning these capacity data, when we look at the difference between the renewable energy capacity (for example solar_capacity and wind_capacity) and the actual generation, we can see that there are large differences. Indeed even during the night the solar_capacity in Great Britain is not 0. This fact first highlights the problems linked to the randomness of this way of producing energy, but also the necessity to work with the actual generation instead of the maximum capacity, indeed the actual generation stays a inferior bound of the capacity.

Another useful information that could be used is the solar_profile or the wind_profile.
For example the term **DE_solar_profile** refers to the solar profile of Germany. Here are the associated details:

- **Type**: Number (represented by a numerical value).
- **Description**: It represents the share of solar capacity produced in Germany. In other words, it indicates the proportion of the total capacity of photovoltaic (PV) solar installations contributing to electricity production in Germany.

In summary, the DE_solar_profile indicates how much solar energy is generated in the German energy mix.

Indeed, the profile could help to palliate the absence of data concerning hydraulic energy which is a problem, since we are concerning byt the sources of renewable energy.

Indeed, here's a breakdown of the energy mix in Switzerland:

- **Hydroelectric power** accounts for about 60% of Swiss-produced energy.
- **Wood** contributes just under 20%.
- Other renewable sources include waste, ambient heat, solar energy, biofuels, biogas, and wind..

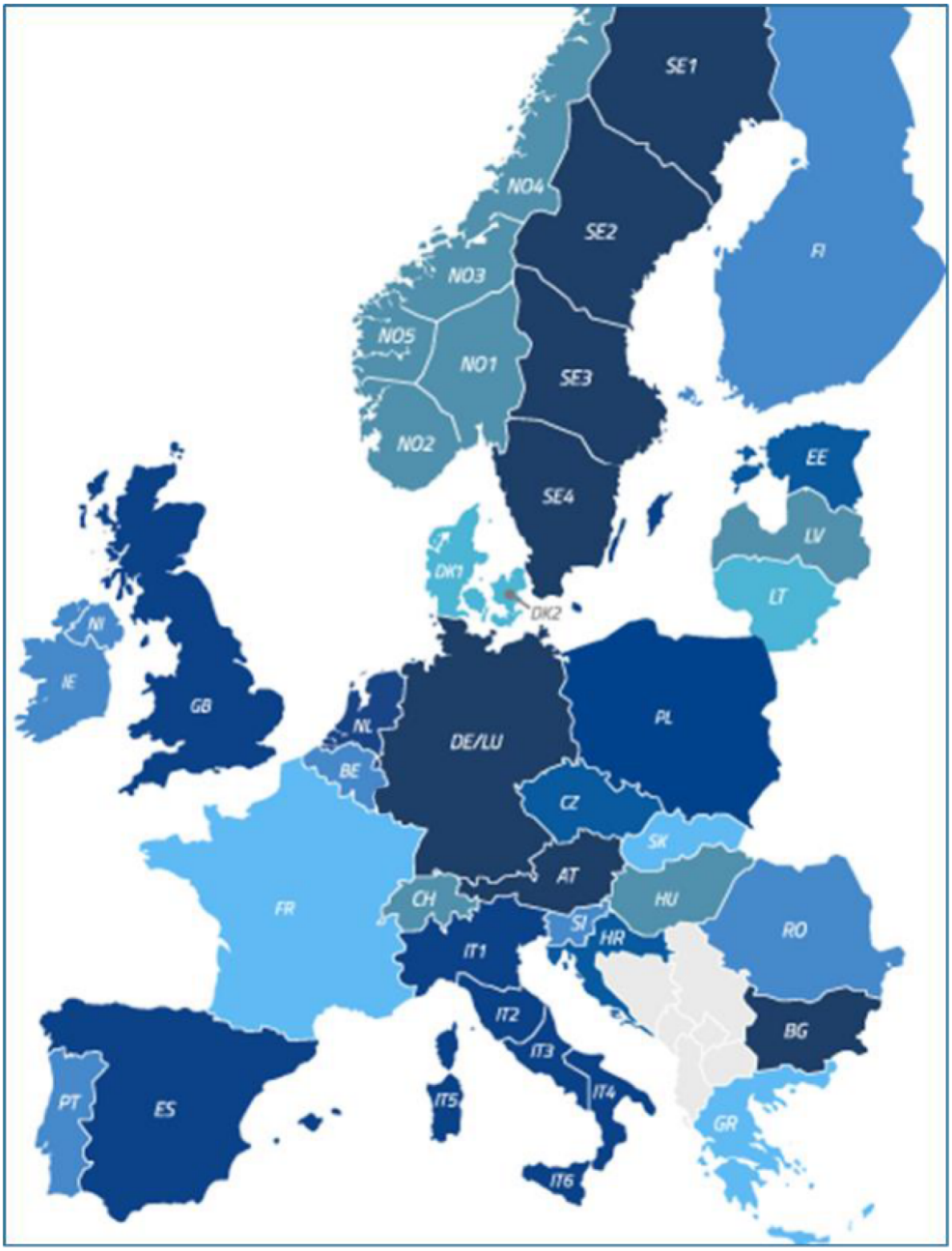
So the hydroelectric power represent a very large part in the energetic mix of Switzerland, so that means more renewable energy could be actually used for the production of hydrogen.

Furthermore, in our study we will keep only one column for the wind energy production. Indeed, the database can present three column for the wind energy production in a country, like presented above for the German case. We can have :

- **Total production**:
  - Refers to wind energy produced on land and generated at sea, it is the mix on onshore and offshore production.
- **Onshore**:
  - Refers to wind energy produced on land, where wind turbines are installed on terrestrial sites.
- **Offshore**:
  - Offshore wind energy is generated at sea, with wind turbines installed offshore, away from the coast.

Here we only keep the column wind_capacity if it is available, otherwise we keep the one available between wind_onshore_capacity and wind_offshore_capacity.
For example, if a country presents only a column called wind_onshore_generation_actual (like Norway), we will change the name into wind_generation_actual.

Here is the map of the bidding zone :




For our study, we will use the following countries :

-Germany (solar_generation_actual and wind_generation_actual)
-Denmark (solar_generation_actual and wind_generation_actual)

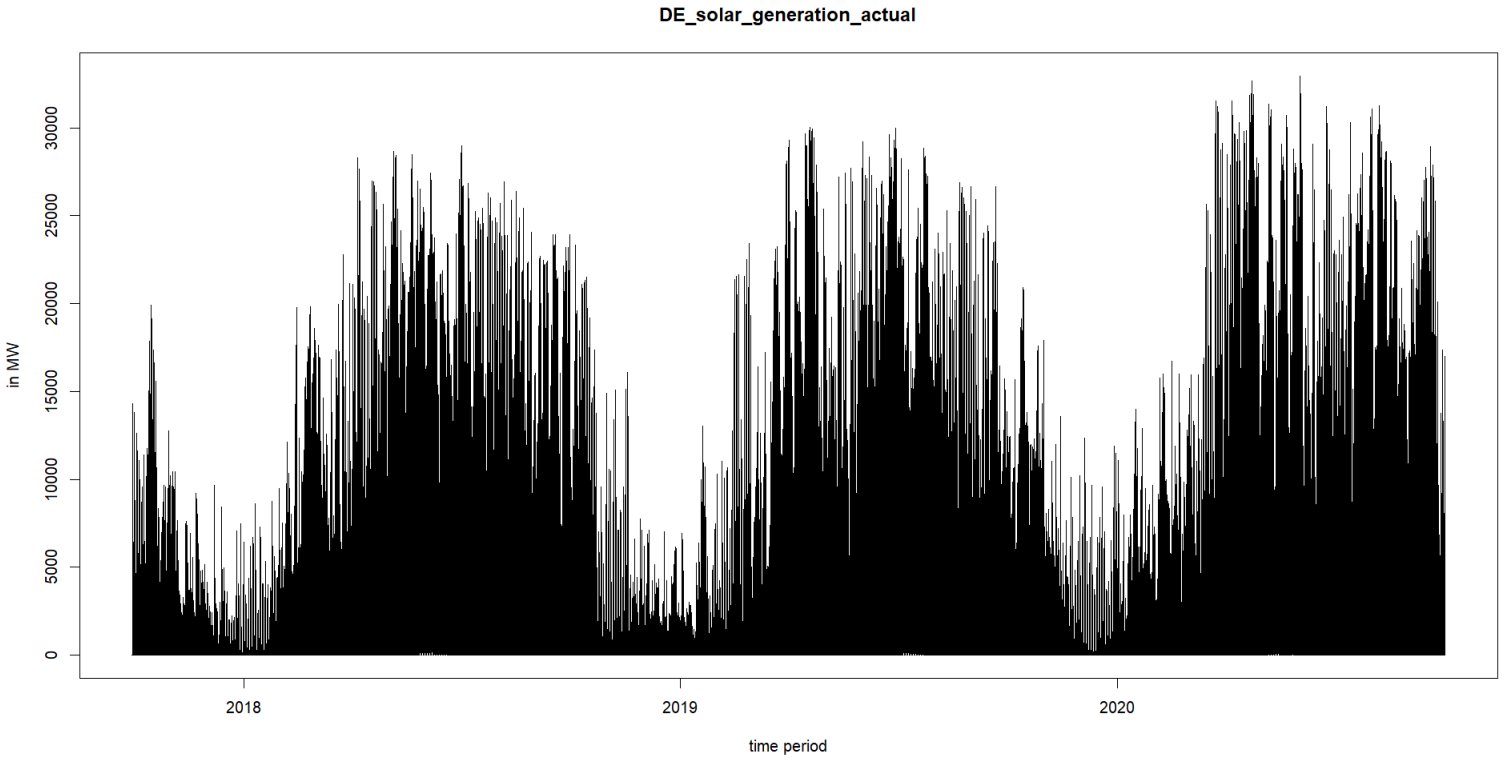
-France (solar_generation_actual and wind_onshore_generation_actual)
-Great Britain (solar_generation_actual and wind_generation_actual)
-Italy (solar_generation_actual and wind_onshore_generation_actual)
-Norway (wind_onshore_generation_actual)
-Sweden (wind_onshore_generation_actual)

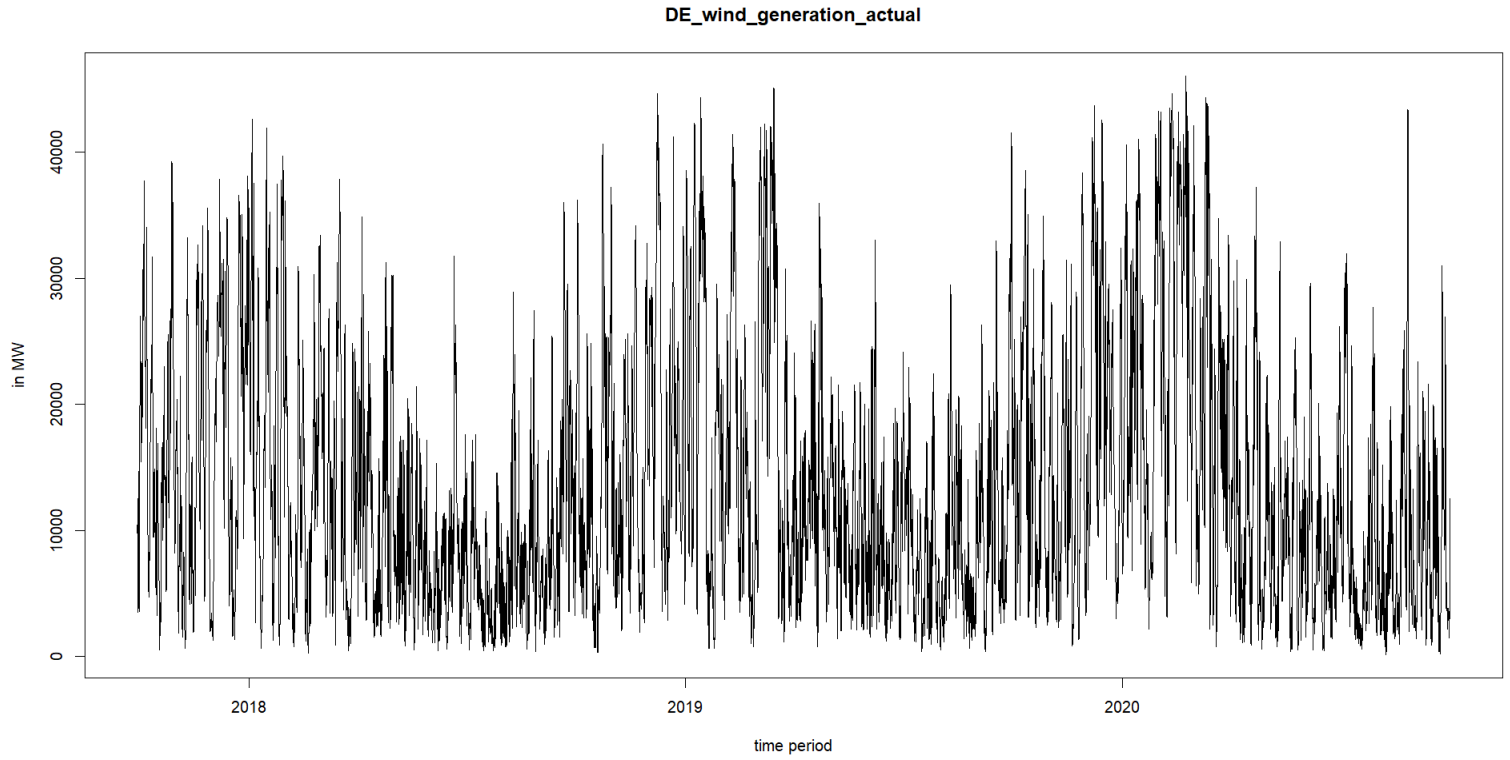
We will use the first bidding zone for Italy, Denmark, Norway and Sweden. It actually does not make a big difference since the prices seem to be the same among the different group for each of these countries.

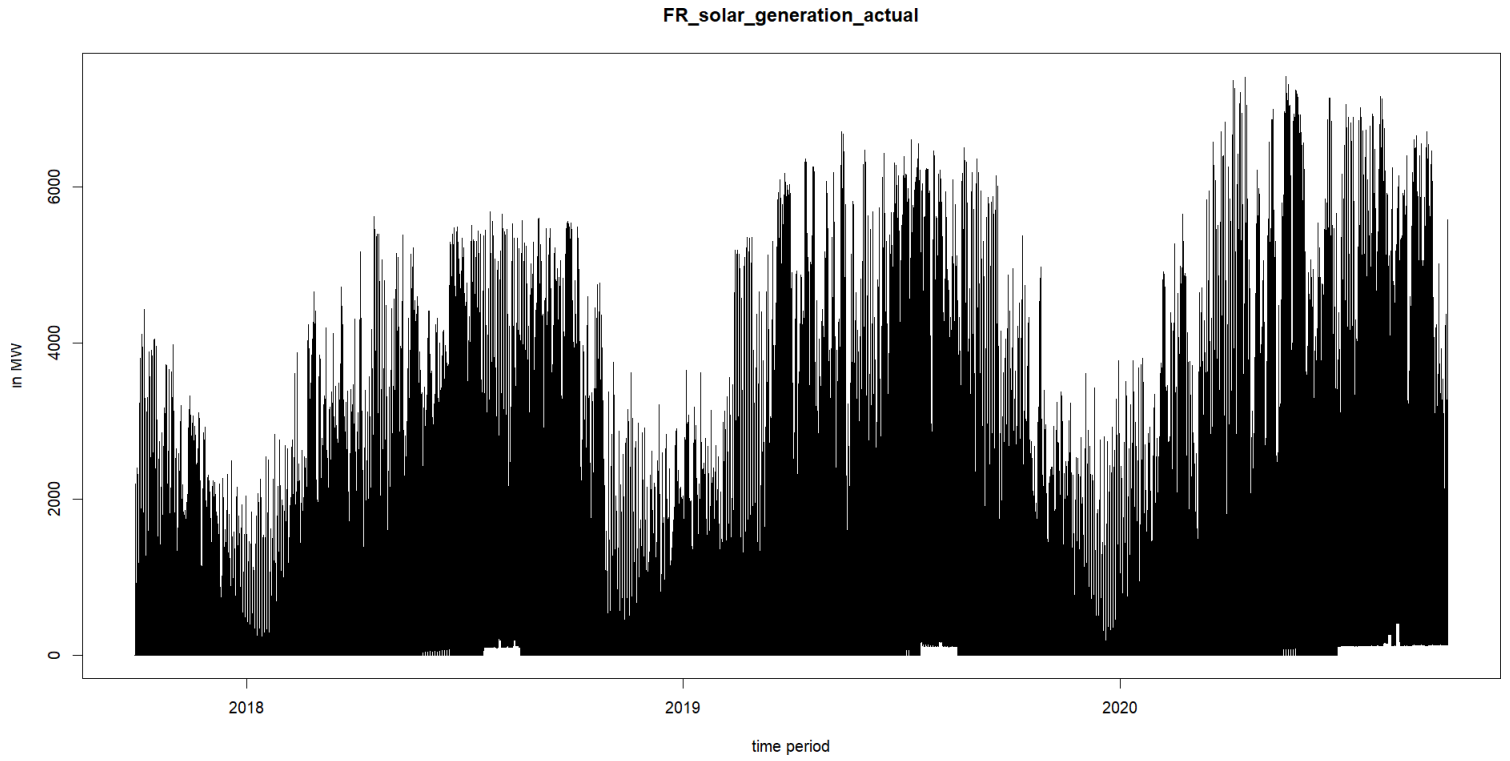
First by observing the selected columns from our initial database after restricting ourselves to the list of chosen countries and the period from 30/09/2017 to 30/09/2020, we can see there are around 180 missing values over 500 232 values, so we have decided to replace these missing values by the average of the concerned column, which is not ideal, but allows to deal with these missing values.

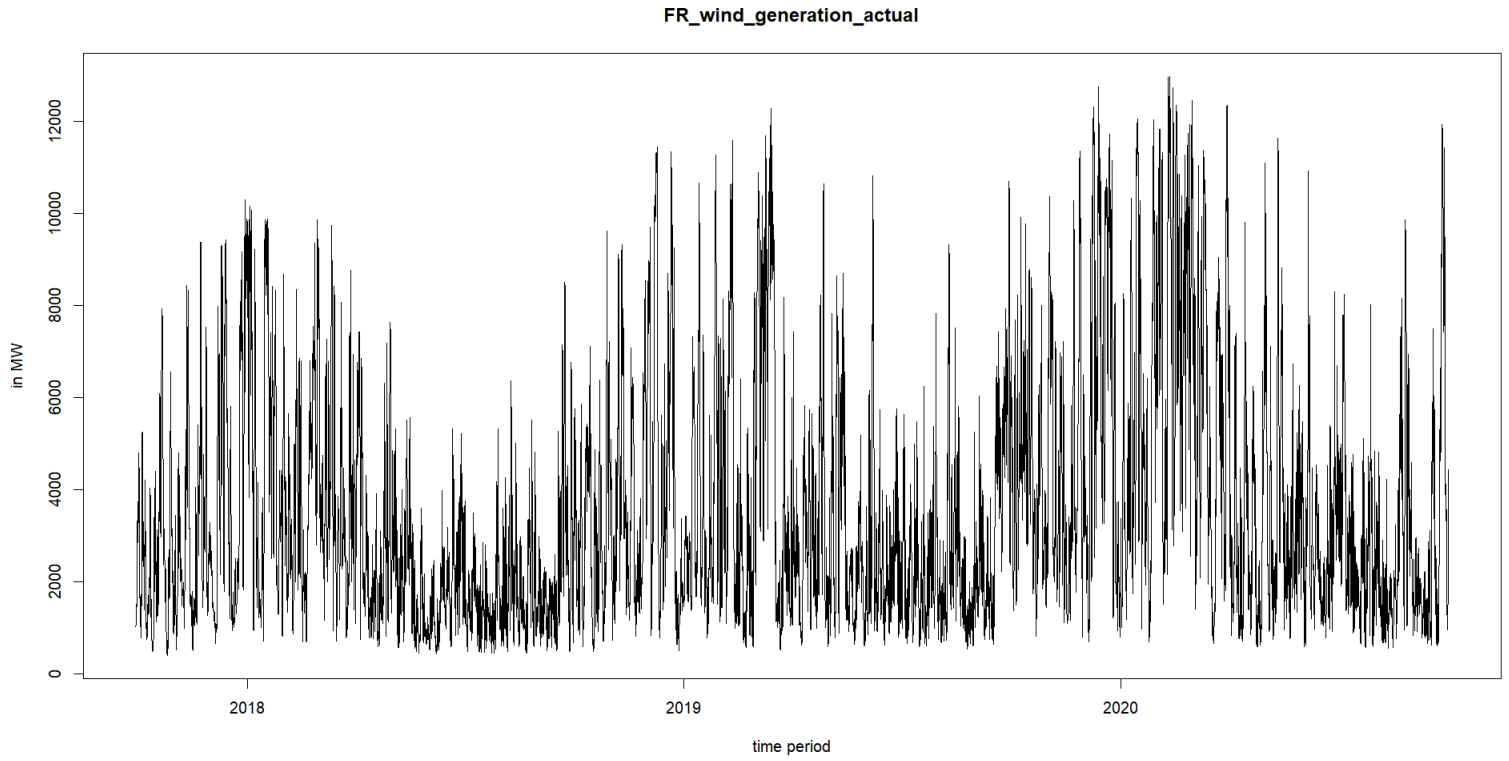
https://www.academia.edu/2767648/Moving_averages

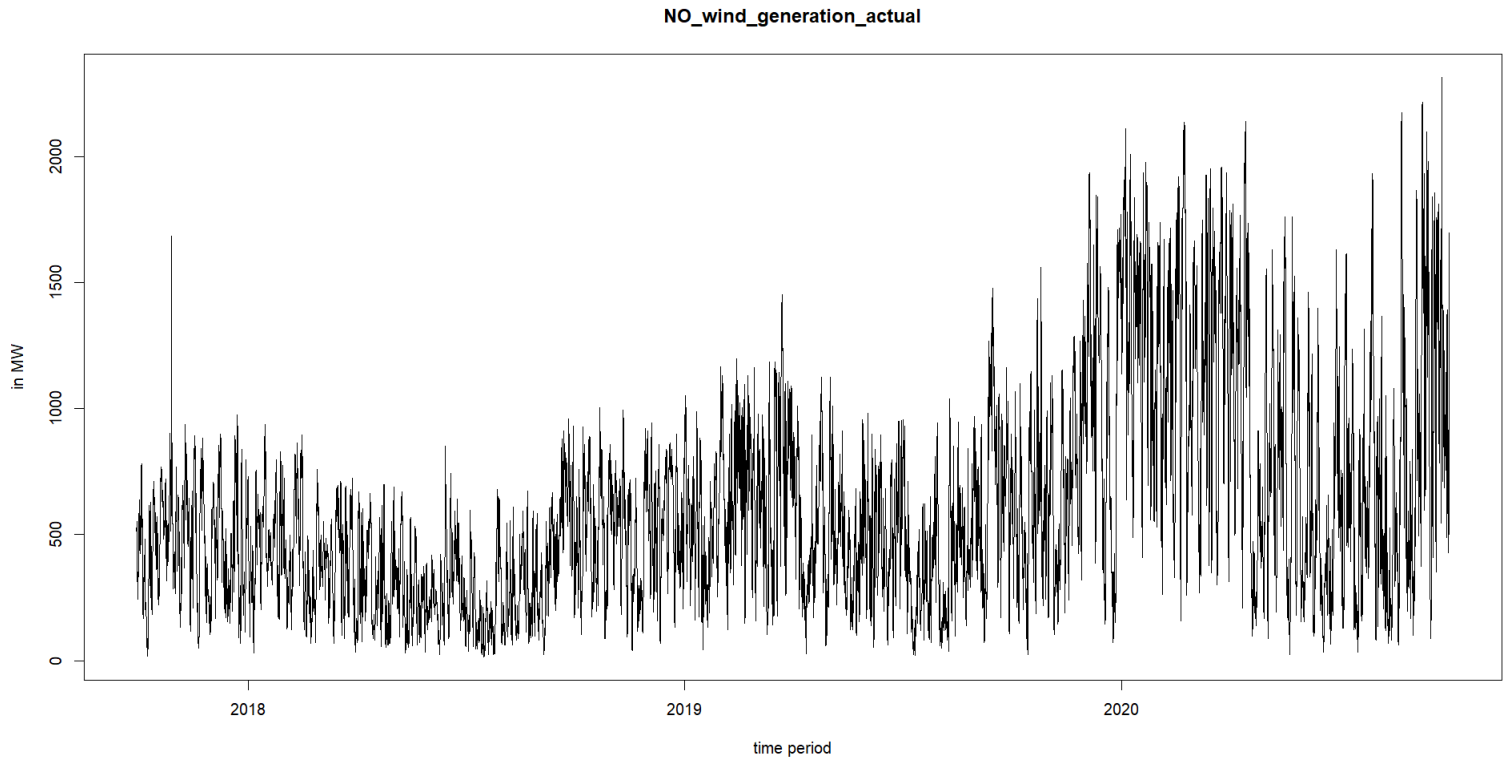
## 2.2 Visualisation of the data




DE_wind_generation_actual
in MW
0
10000
20000
30000
40000
2018
2019
2020
time period

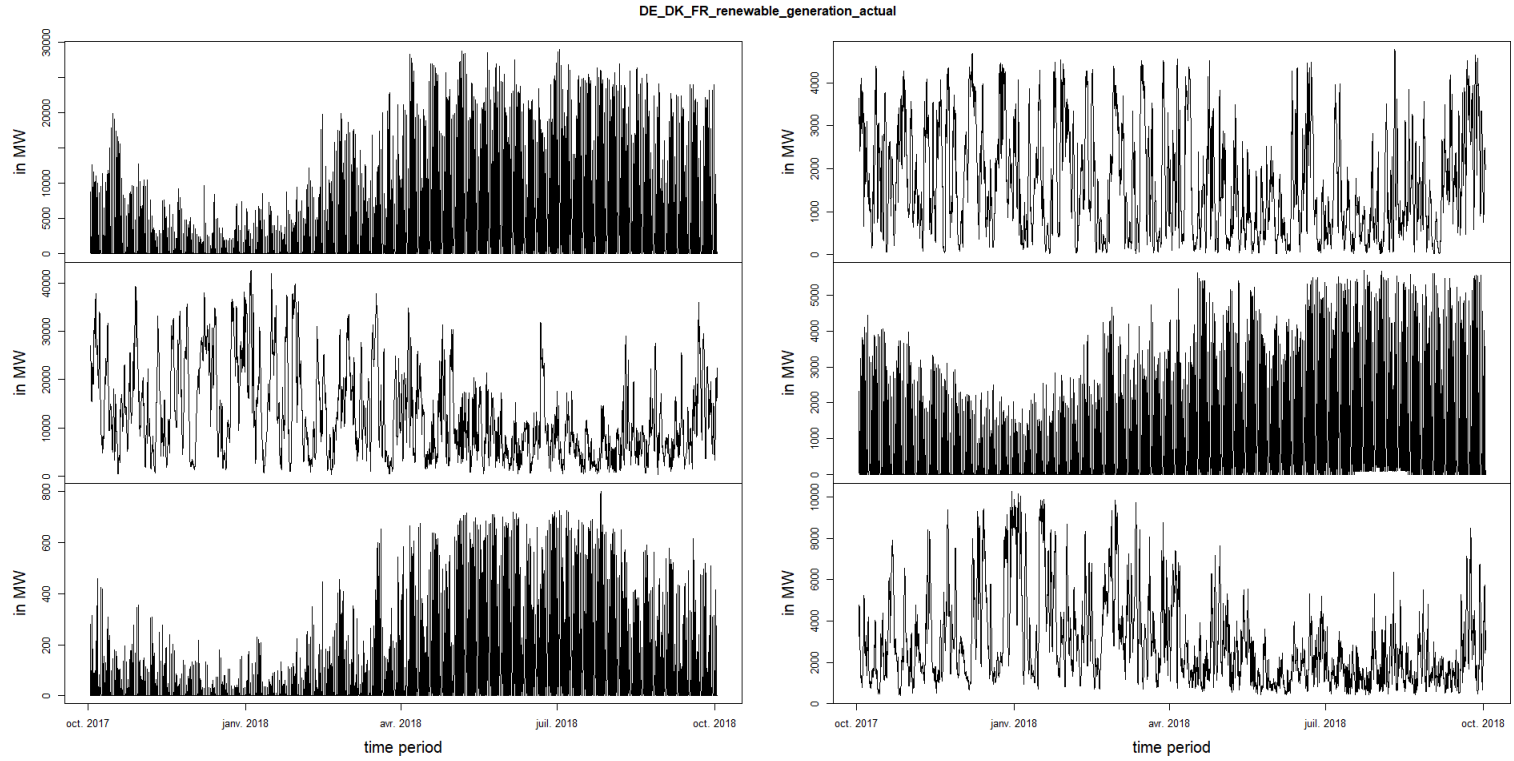

FR_solar_generation_actual
in MW
0
2000
4000
6000
2018
2019
2020
time period



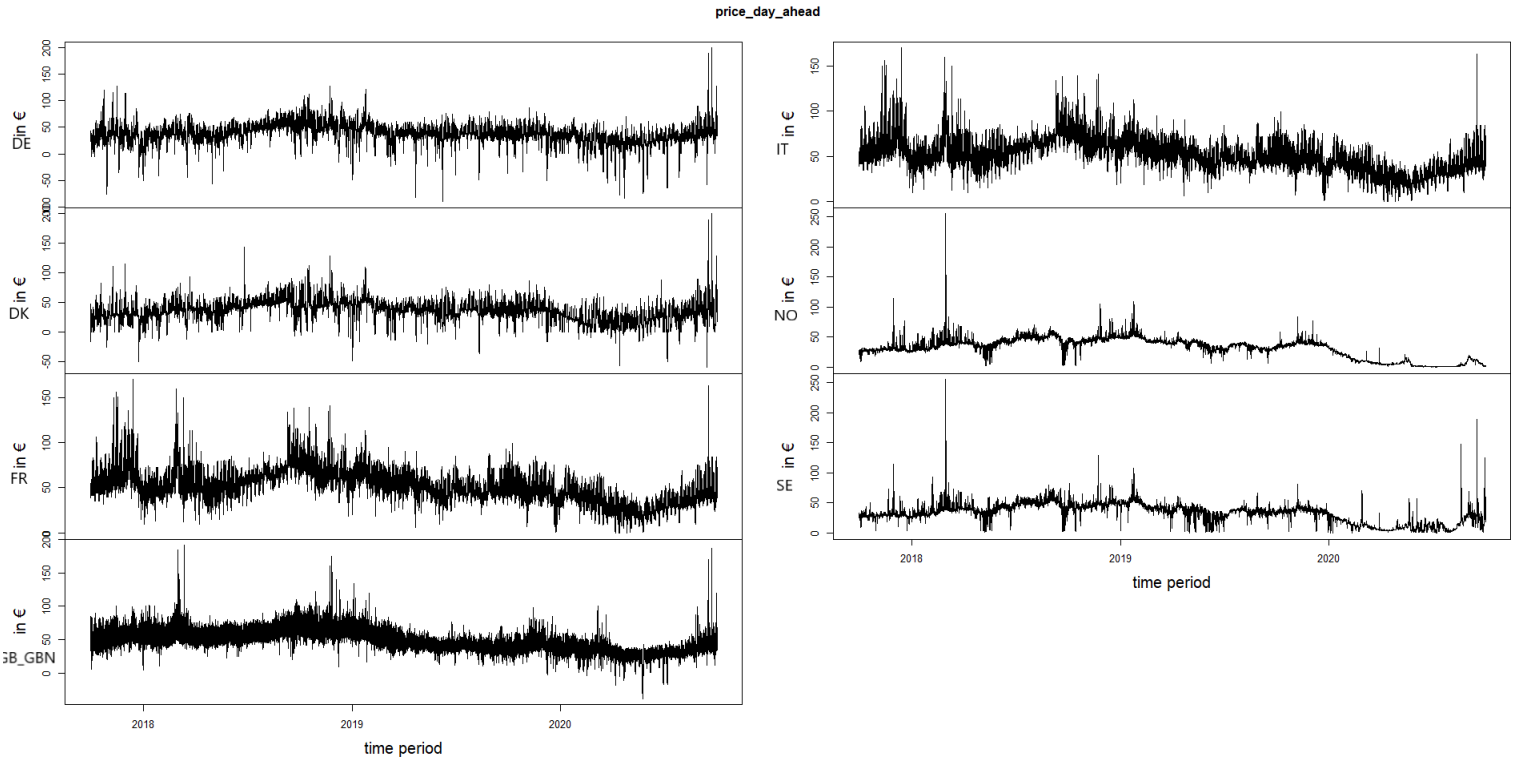




We can see with this first visualization some trends, indeed during the summer the production of electricity is higher than during winter, and this does not change over the years. For the wind energy production, it seems that there is more wind at the beginning of the year around January, nevertheless

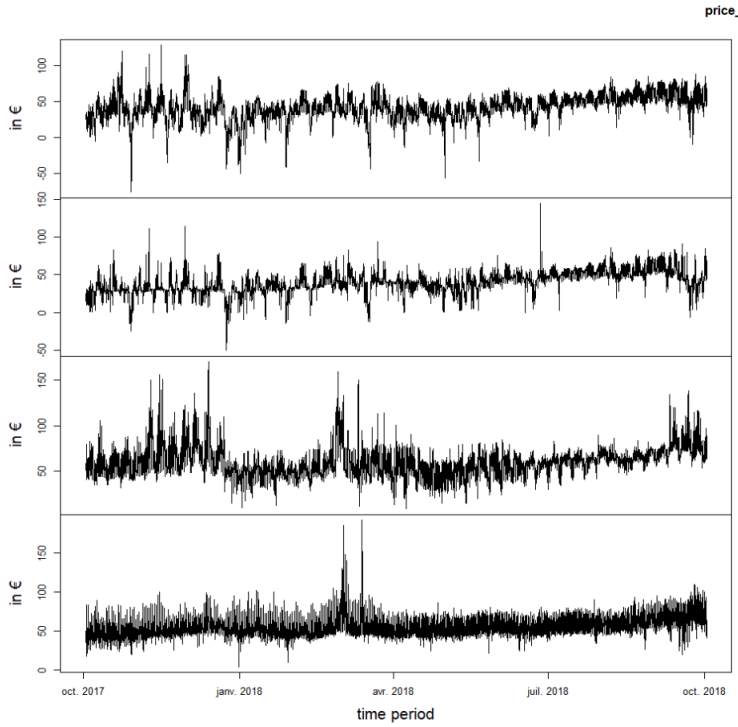
this production seems more random than the solar energy production.

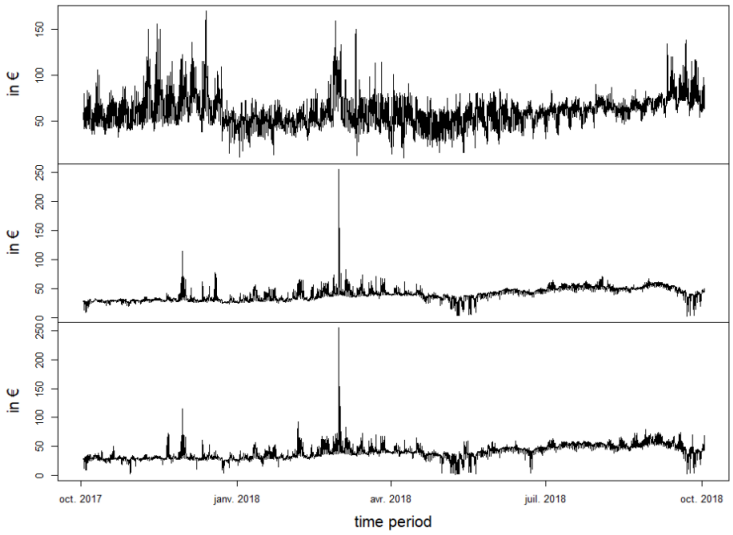


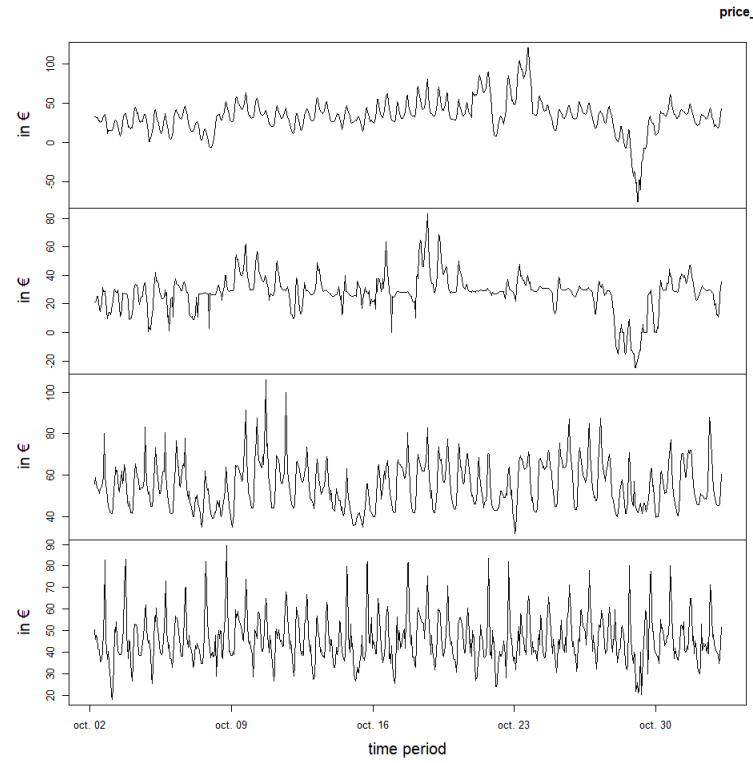

This is a more condensed graphic, where the period is actually restricted to one year from 30/10/2017 to 30/10/2018, this allows to highlight the previous trends seen previously.

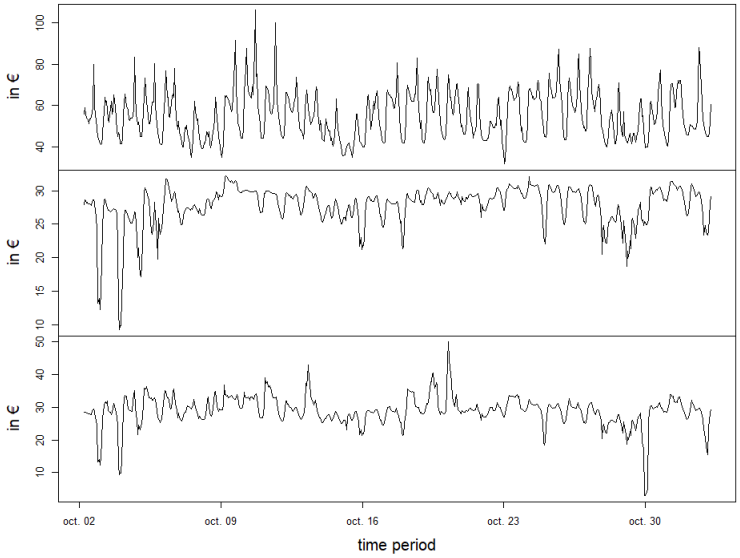


We can see that the price_day_ahead among the selected European countries do not seem to present any trend, except that it seems that these prices among European countries present similar growth.



When we restrict the time period, to one year and then to one year, we see that the previous observations still seem to be accurate.

## 2.3 Statistical analysis

This section aims at giving some statistical evidences of the previous observations using time series analysis background.

The source used for this part is : https://www.statslab.cam.ac.uk/%7Errw1/timeseries/t.pdf

One simple method of describing a series is that of classical decomposition. The notion is that the series can be decomposed into four elements:

- **Trend** ($T_t$) — long-term movements in the mean.
- **Seasonal effects** ($I_t$) — cyclical fluctuations related to the calendar.
- **Cycles** ($C_t$) — other cyclical fluctuations (such as business cycles).
- **Residuals** ($E_t$) — other random or systematic fluctuations.

The idea is to create separate models for these four elements and then combine them, either additively:

$$X_t = T_t + I_t + C_t + E_t$$

or multiplicatively:

$$X_t = T_t \cdot I_t \cdot C_t \cdot E_t$$

1. A sequence $\{X_t, t \in \mathbb{Z}\}$ is **strongly stationary** or **strictly stationary** if

$$(X_{t_1}, \ldots, X_{t_k}) \stackrel{D}{=} (X_{t_1+h}, \ldots, X_{t_k+h})$$

for all sets of time points $t_1, \ldots, t_k$ and integer $h$.

2. A sequence is **weakly stationary**, or **second-order stationary**, if:
- (a) $\mathbb{E}(X_t) = \mu$, and
- (b) $\text{cov}(X_t, X_{t+k}) = \gamma_k$, where $\mu$ is constant and $\gamma_k$ is independent of $t$.

3. The sequence $\{\gamma_k, k \in \mathbb{Z}\}$ is called the **autocovariance function**.

4. We also define

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \text{corr}(X_t, X_{t+k})$$

and call $\{\rho_k, k \in \mathbb{Z}\}$ the **autocorrelation function (ACF)**.

**Analysis for the price_day_ahead in Germany**

To begin with, we are actually going to apply these concept to the price_day_ahead time series of Germany.

However, a very simple diagnostic is the **turning point test**, which examines a series $\{X_t\}$ to test whether it is purely random. The idea is that if $\{X_t\}$ is purely random then three successive values are equally likely to occur in any of the six possible orders.

In four cases there is a turning point in the middle. Thus in a series of $n$ points we might expect $(2/3)(n-2)$ turning points.

In fact, it can be shown that for large $n$, the number of turning points should be distributed as about $N(2n/3, 8n/45)$. We reject (at the 5% level) the hypothesis that the series is unsystematic if the number of turning points lies outside the range $2n/3 \pm 1.96\sqrt{8n/45}$.

So in order to see if the price_day_ahead of Germany is as random as we thought we run a turning point test. Here is the result :

```
        Turning point test of independence

data:  DE_price_ahead_2017_2020
T = -250.96, p-value < 2.2e-16
```

We actually observe a very small p-value, which means that we reject the null hypothesis, and therefore we consider that this time series presents some structure.

This actually makes sense, indeed from one hour to the other the price will not be drastically different, it can increase or decrease but not change completely.

Nevertheless, even if this time series presents some structure, we can test whether it actually presents some patterns or some trends. Indeed, our previous observation concerning this price time series was too strong, although the prices from one hour to another are not completely random, that does not

mean this times series present some trends or patterns over times.

So what we want actually test is whether this time series is stationary, to do so we apply an Augmented Dickey Fuller (ADF) test. Here is the result :

```
> adf.test(DE_price_ahead_2017_2020)

        Augmented Dickey-Fuller Test

data:  DE_price_ahead_2017_2020
Dickey-Fuller = -17.764, Lag order = 29, p-value = 0.01
alternative hypothesis: stationary
```

Since the p-value is small enough (less than 0.05), we can reject the null hypothesis, and therefore ensure that this time series is stationary.

To finalize our analysis of this stationary time series, we compute an ARIMA model, which allows to predict future prices.

Here is the summary of this model :

```
Series: DE_price_ahead_2017_2020
ARIMA(2,1,2)

Coefficients:
         ar1      ar2      ma1     ma2
      1.4579  -0.6325  -1.0736  0.1655
s.e.  0.0094   0.0098   0.0119  0.0127

sigma^2 = 20.95:  log likelihood = -77240.21
AIC=154490.4   AICc=154490.4   BIC=154531.3

Training set error measures:
                      ME     RMSE      MAE MPE MAPE       MASE          ACF1
Training set 0.001214086 4.576503 2.893255 NaN  Inf 0.07604194 0.0006525555
```

The numbers within the parentheses represent the model's parameters:

- **AR (Auto-Regressive) Terms** ($p = 2$):
  - These terms capture the linear relationship between the current value and its previous values (lags).
  - In our model, we have two AR terms. Each term corresponds to a specific lag (e.g., AR(1) and AR(2)).

- A higher value of $p$ indicates that the model considers more historical observations for prediction.
- If the AR coefficients are significant, it suggests that the time series depends on its own past values.

- **Differencing Order** ($d = 1$):
  - The differencing order determines how many times we need to difference the series to make it stationary.
  - First-order differencing ($d = 1$) means subtracting each value from its previous value.
  - Stationarity is crucial because many time series models assume that the data is stationary (i.e., constant mean and variance).

- **MA (Moving Average) Terms** ($q = 2$):
  - These terms account for the relationship between the current value and past forecast errors (residuals).
  - In our model, we have two MA terms (e.g., MA(1) and MA(2)).
  - A higher value of $q$ indicates that the model considers more past forecast errors for prediction.
  - Significant MA coefficients imply that the model adjusts predictions based on past errors.

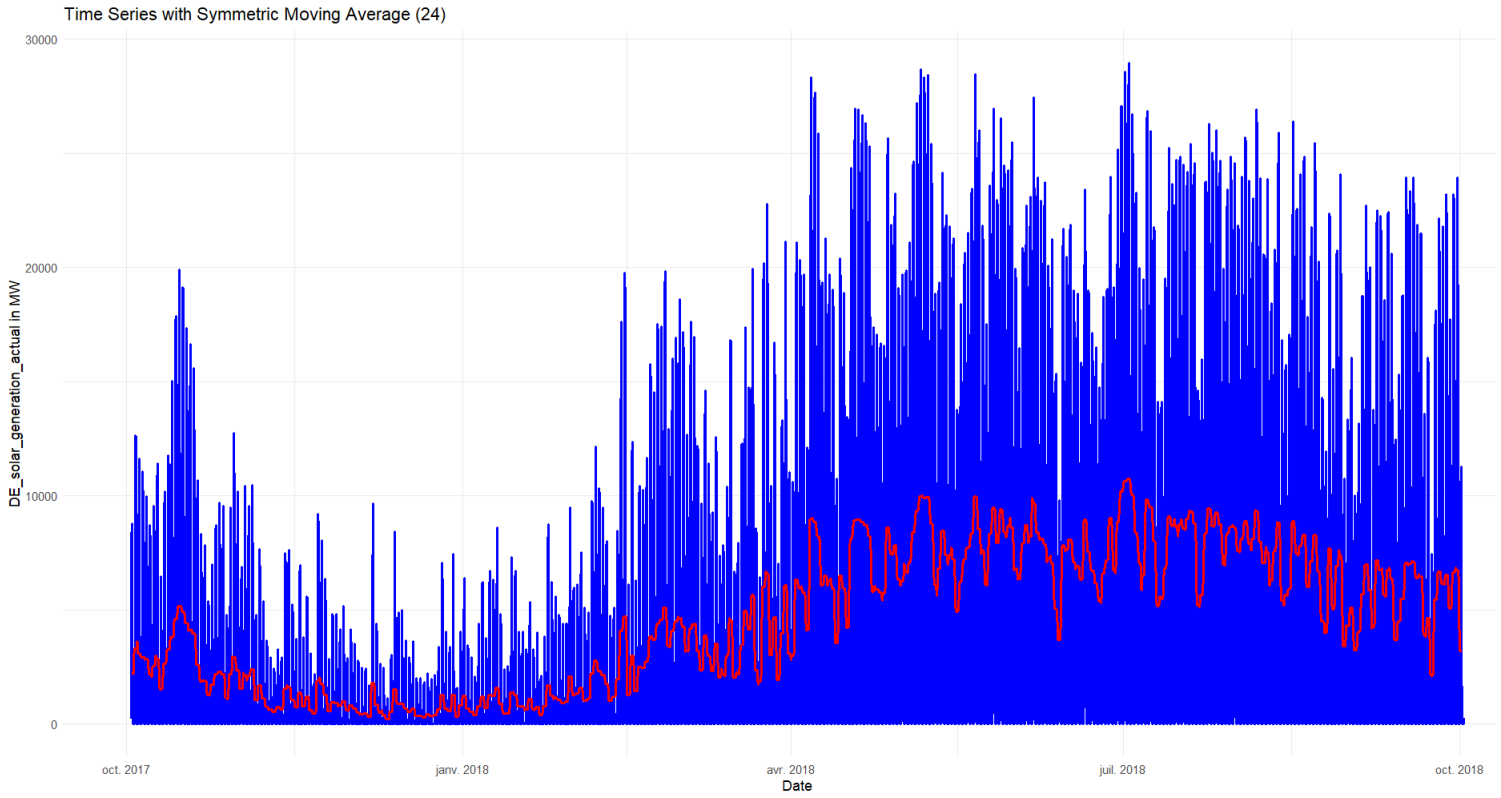
An interesting point is that the model says that we would actually need to differentiate once to get a stationary time series, this difference could come from the threshold chosen to reject the null hypothesis in the ADF test. But the other values p ad q are quite small, s as it is explained, we can conclude from this model that this time series actually presents a simple structure, only the recent prices seem to have an impact on the new one, but there is no trend or pattern.

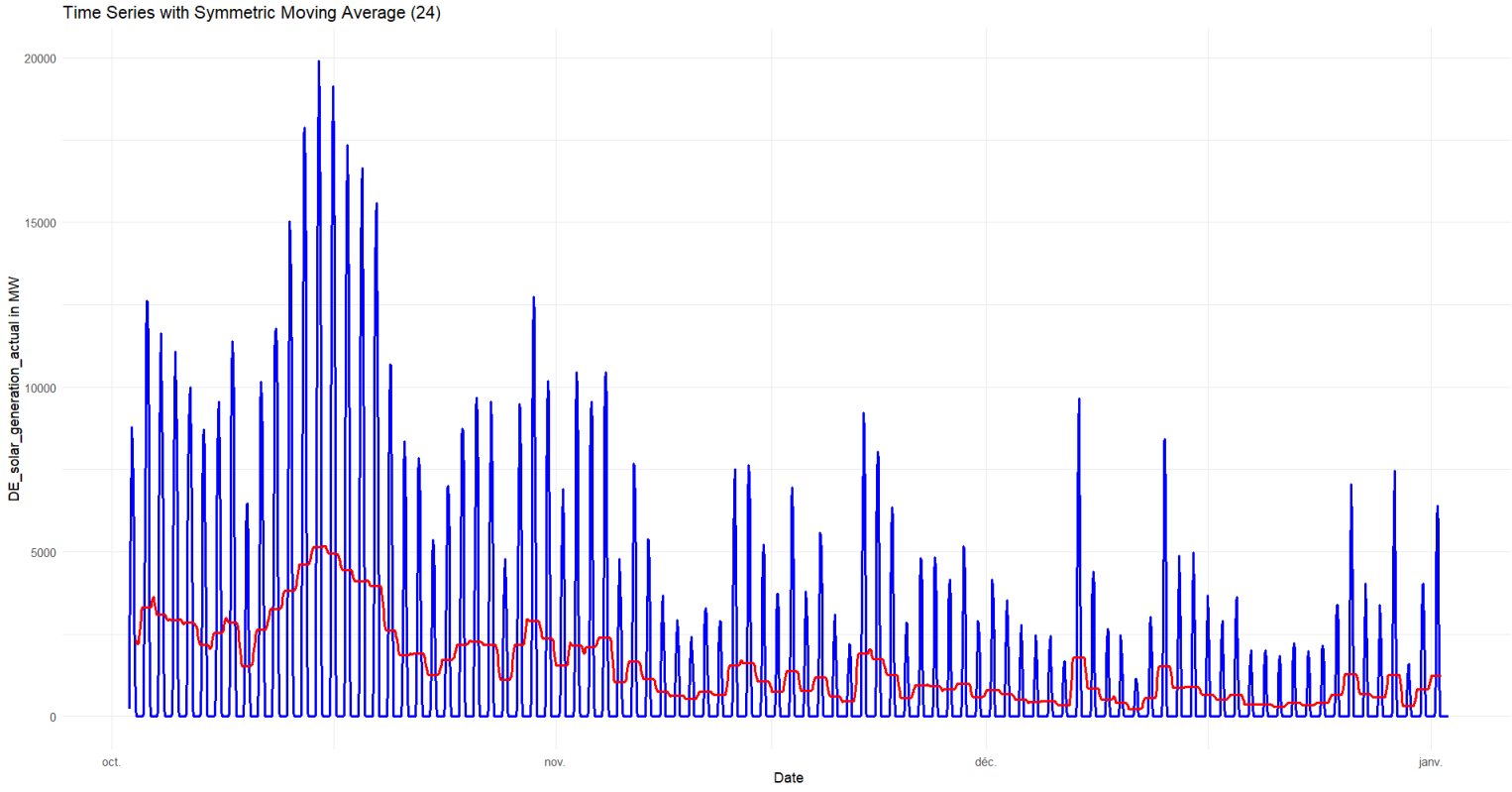
**Analysis for the renewable energy actual generation in Germany**

To run our analysis we will focus on the time period going from October 2017 to October 2018.

**For the solar_generation_actual**

Here, we plot the data and we also add the symmetric moving average computed with a lag of 24. The idea is to average the data per day, in order to smooth the curve.




Now, we are going to do as previously, that means we are going to look for the presence of trends or patterns. To do so we run a ADF test, here is the result :

```
> adf.test(DE_solar_generation_actual_2017_2018)

        Augmented Dickey-Fuller Test

data:  DE_solar_generation_actual_2017_2018
Dickey-Fuller = -3.8095, Lag order = 20, p-value = 0.01844
alternative hypothesis: stationary
```

This is interesting, we can first see that the p-value is higher than when we ran the same test over the price time series, but it still quite low, although we can see summer pics or even pick during the day.

Therefore, to study better the presence of trends we are going to plot an ADF and PACF graph.

**Autocorrelation Function (ACF):** The ACF measures the correlation between a time series and its lagged values (previous observations). It helps identify overall patterns in the time series. Specifically, the ACF at lag $k$ represents the correlation between the series at time $t$ and the series at time $t - k$. ACF values range from -1 to 1:

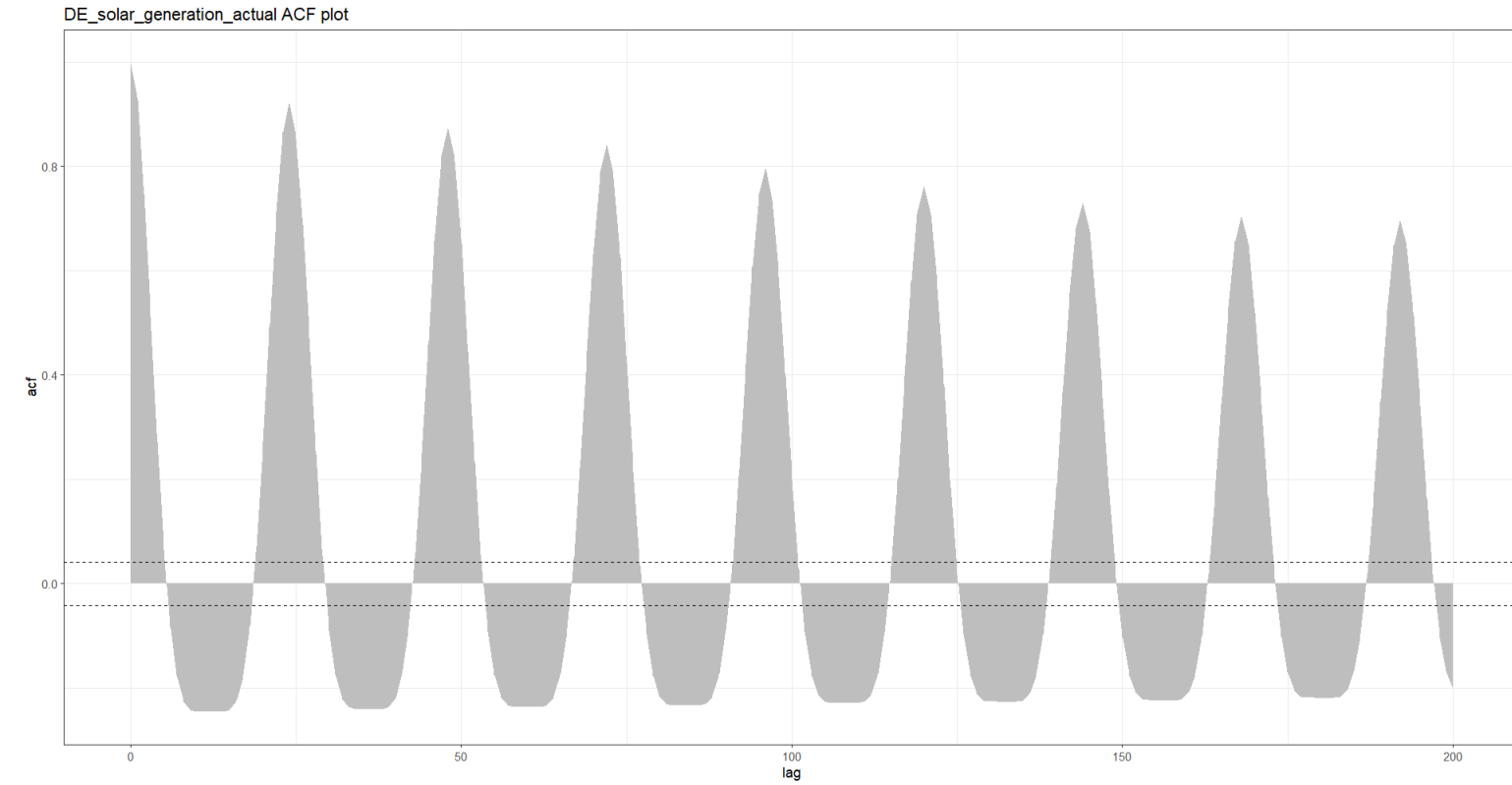
- **Positive ACF**: Indicates positive correlation (similar patterns repeating over time).
- **Negative ACF**: Indicates negative correlation (opposite patterns).
- **ACF near zero**: Suggests no significant correlation.

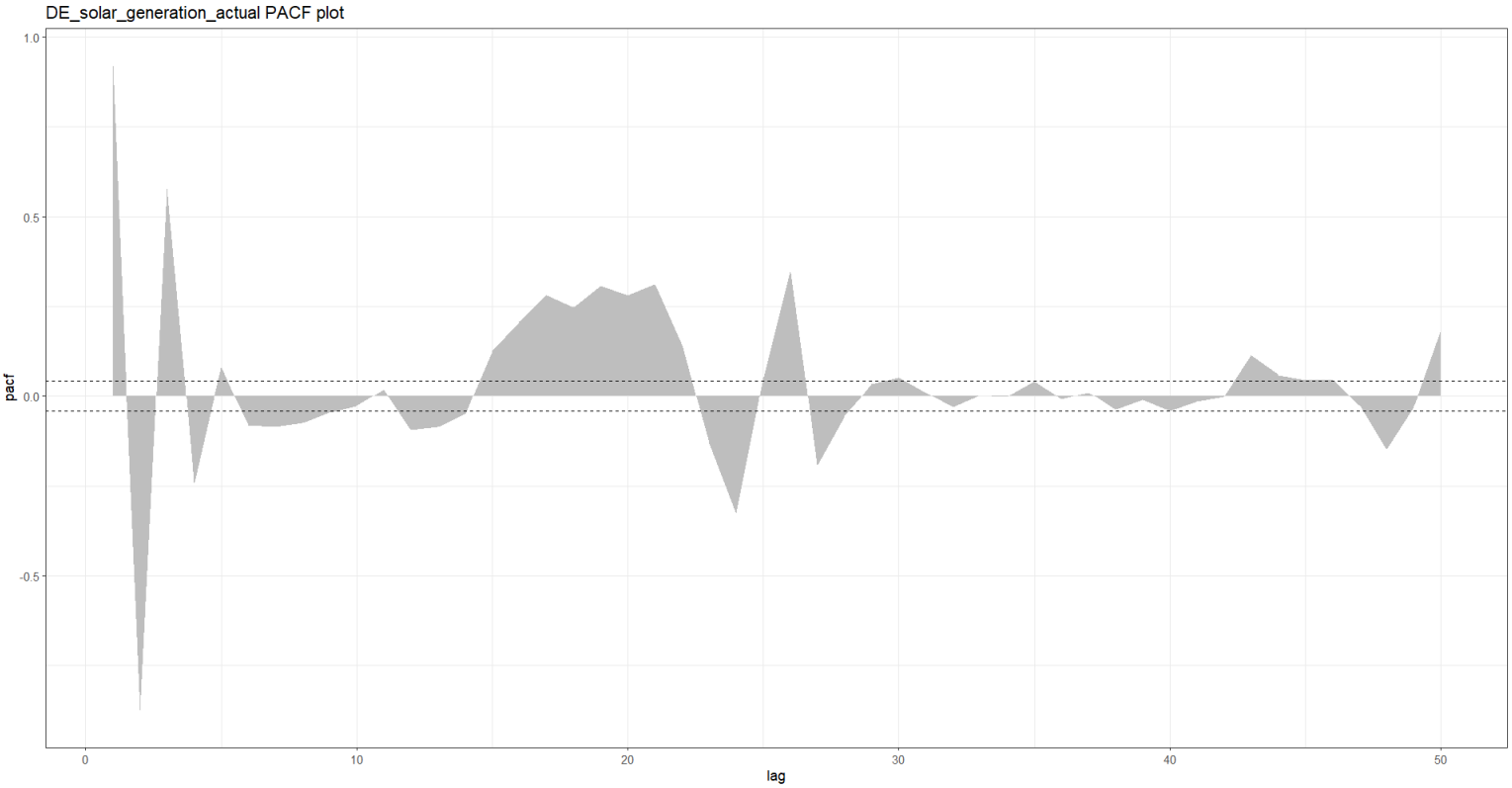
**Partial Autocorrelation Function (PACF):** The PACF refines the analysis by highlighting direct correlations between a time series and its lagged values, while removing indirect correlations. It measures the correlation between a value and its lag, controlling for the influence of intermediate lags. Specifically, the PACF at lag $k$ represents the correlation between the series at time $t$ and the series at time $t - k$, excluding the influence of lags in between. PACF values also range from -1 to 1:

- **Positive PACF**: Indicates a direct correlation after removing the effect of intermediate lags.
- **Negative PACF**: Indicates a direct negative correlation.
- **PACF near zero**: Suggests no significant direct correlation.

**Interpretation:** By comparing the two plots (ACF and PACF), we can observe how they complement each other in revealing the temporal dependencies within the time series. The ACF helps identify overall patterns, while the PACF provides more specific information about direct correlations.

Here are the plots :



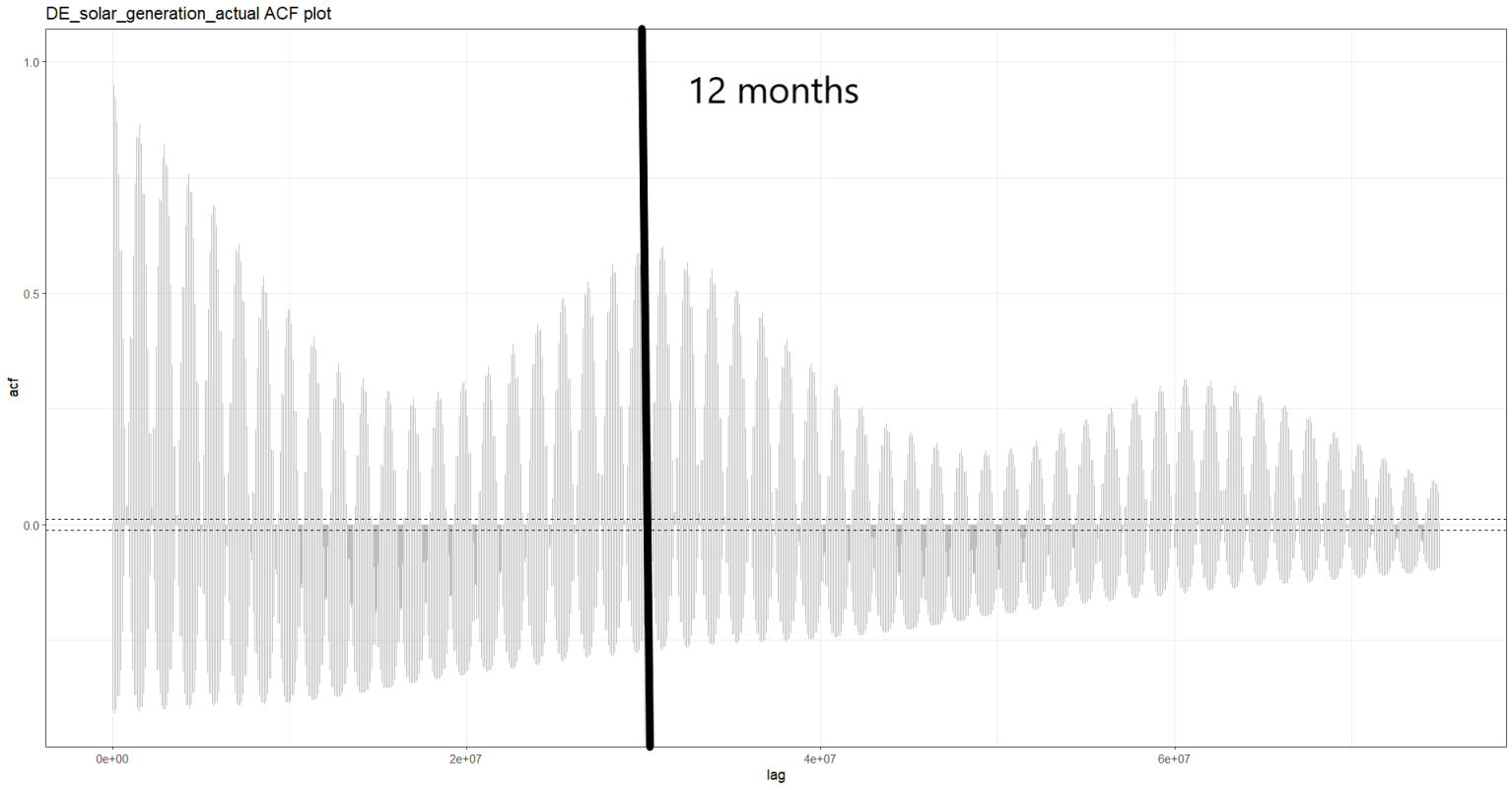



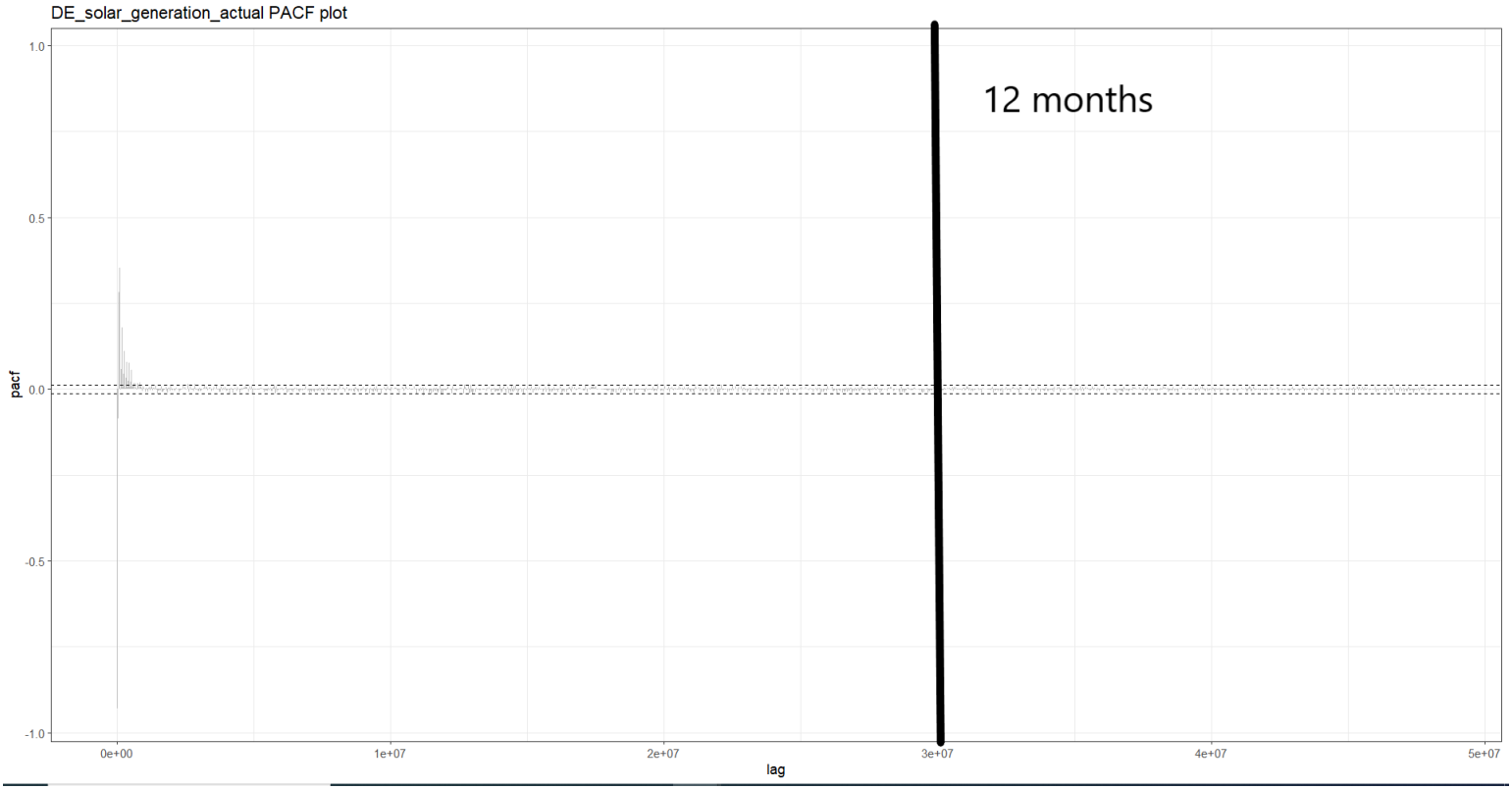

The first ADF plot is quite interesting. Indeed, we can clearly see some patterns : at every lag

multiple of 24 we see a positive peak, which means that from one day to another the evolution are the same, that means in our case that the solar energy production increases in the first part of the day around 7 am and then decreases in the afternoon, as we can see on the first graph of this part.
The other interesting point is the presence of negative points at lag that are odd multiple of 12, which is realist since at some point it is the middle of the night and then twelve hours later it is the zenith (to summarize).
Moreover the size of these picks decrease with the number of lags, that means that the impact added by the previous hours decrease with the number of hours. Indeed, we have daily cycle, so the data of the previous hours are enough to predict what could be the values of the following days.

Concerning the PACF plot we see something very interesting. Indeed compared to the ACF plot the impact of the previous values are taken independently, so instead of looking at the impact of the k latest value on the new value, we just look at the impact of the k-value on the new value. This allows to get rid of the natural trend, which considers in having close value in terms of solar energy production for close hours, but it allows to highlight the impact of precise time differences. So here we can see that the lags around 23 have significant impact, which is coherent again with the reality. Therefore we can ensure the presence of daily cycles.

Now we want to look at the presence of monthly cycles such as seasonal trends, like the fact that we have more sun in the summer than during the winter. To do so we are going to use the time period from 30/09/2017 to 30/09/2020.






So here we can see with the ACF plot what we previously observed the presence of an annual pattern, but the PACF does not confirm this observations, this can come from the fact that we have only three years which is not enough to highlight an annual trend.

Now we are also going to run an ARIMA model on our time series. First we restrict of the period from October 2017 to October 2018. Here is the summary of the model :

```
> summary(arima_model)
Series: DE_solar_generation_actual_2017_2018
ARIMA(3,1,2) with drift

Coefficients:
         ar1     ar2      ar3     ma1    ma2    drift
      0.8522  0.2007  -0.3851  1.0686  0.303  -0.0563
s.e.  0.1306  0.1931   0.0901  0.1291  0.056  40.0519

sigma^2 = 276037:  log likelihood = -67302.83
AIC=134619.7   AICc=134619.7   BIC=134669.2

Training set error measures:
                      ME     RMSE      MAE MPE MAPE      MASE         ACF1
Training set 0.008738535 525.1823 322.3832 NaN  Inf 0.0701994 -0.00716269
```

What is interesting with this ARIMA model is the fact that the values of the model are quite low, which seems to indicate that the most recent value are enough to explain the new data. Indeed we are actually in front of a usual trading in statistics, the fact of having the simplest model which is the best o explain the data. So there is no need to look at the previous days if looking at the three past hours is enough to have already a pretty strong forecasting power.

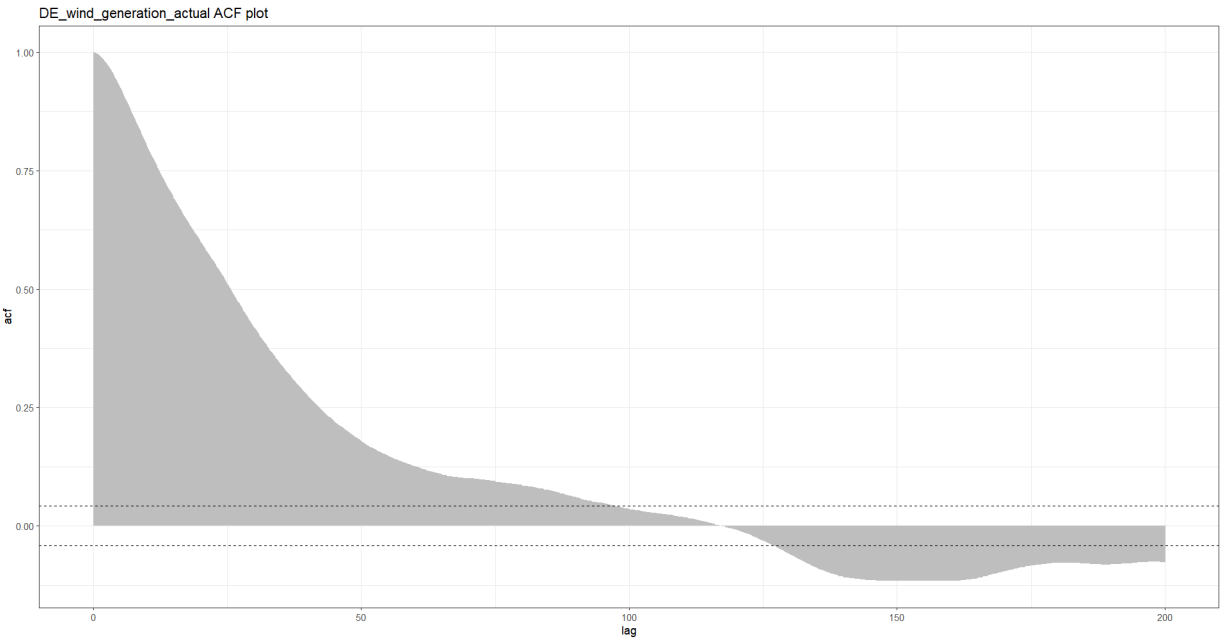
Nevertheless, we are construct our own ARIMA model, so constraint some parameters to use the daily patter that we have previously observed. Here is the summary of the model :

```
> summary(sarima_model)
Series: DE_solar_generation_actual_2017_2018
ARIMA(1,1,1)(1,1,1)[24]

Coefficients:
         ar1     ma1    sar1     sma1
      0.6675  0.5871  0.1440  -0.8220
s.e.  0.0086  0.0080  0.0135   0.0073

sigma^2 = 88791:  log likelihood = -62174.58
AIC=124359.2   AICc=124359.2   BIC=124394.5

Training set error measures:
                      ME     RMSE      MAE MPE MAPE      MASE      ACF1
Training set -0.09517395 297.4843 178.3616 NaN  Inf 0.0388385 0.1021991
```

We can first see that the RMSE and MAE value are lower for this model than for the previous model, which indicates better performance of the SARIMA model.

```
> # Compare accuracy metrics (e.g., MAE, RMSE, MAPE)
> accuracy(arima_forecast)
                     ME   RMSE      MAE MPE MAPE      MASE       ACF1
Training set 0.07826716 401.17 224.8535 NaN  Inf 0.1462038 0.02300068
> accuracy(sarima_forecast)
                      ME     RMSE      MAE MPE MAPE       MASE       ACF1
Training set -0.1939725 247.1511 132.8878 NaN  Inf 0.08640605 0.08972973
```

Now for the wind energy production of Germany, we plot directly the ADF plot in order to see if the previous observation, which consisted in saying that the time series did not present pattern is significant :

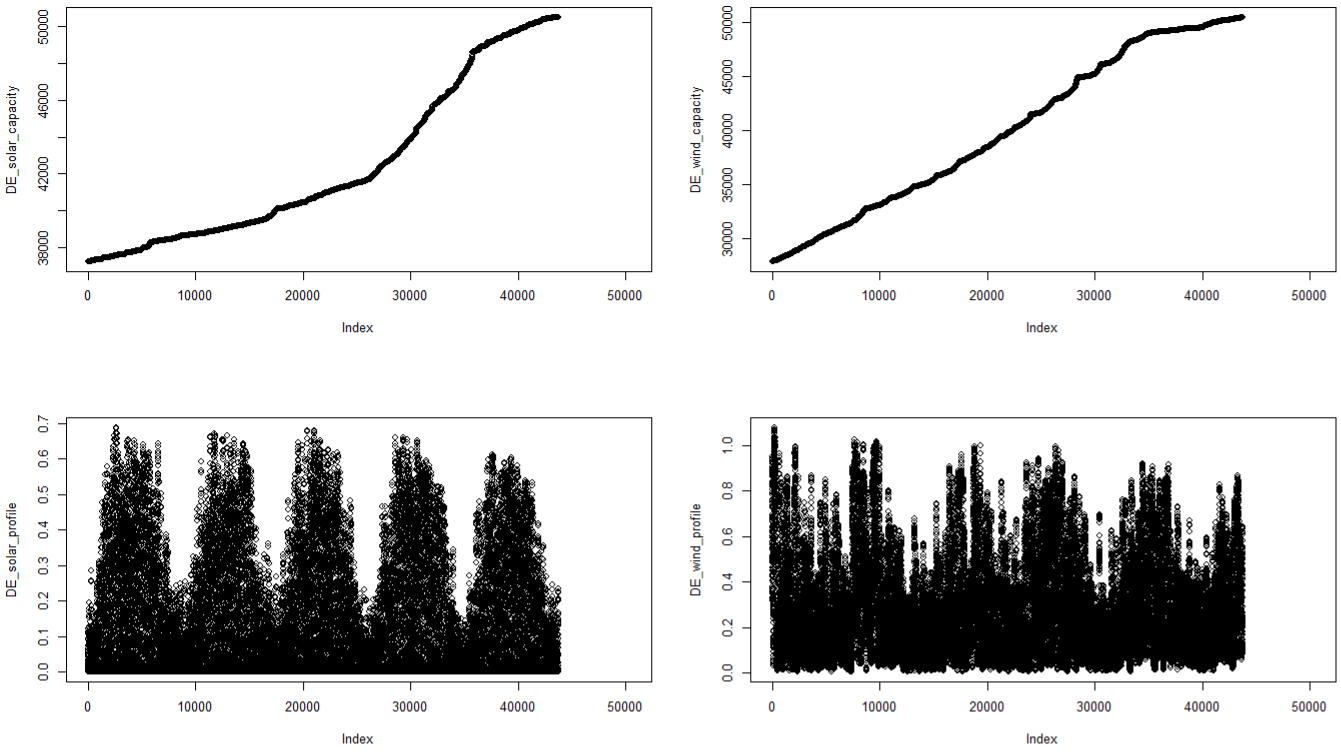



The fact that we have a decreasing slope for the ADF plot comforts in the idea that the wind energy values are not completely random, this means that data close in terms of hours are also close in terms of values, but we clearly do not have a pattern.

# 3 Selection of data

## 3.1 Choice of the columns

Here we will work with the profile of energy and not the actual generation of energy. Indeed, we want to work on a fixed entity, that means we have our solar panels and wind turbines, so we will not increase these capacities over the weeks and the years. To be more precised we are actually going to focus on Germany.



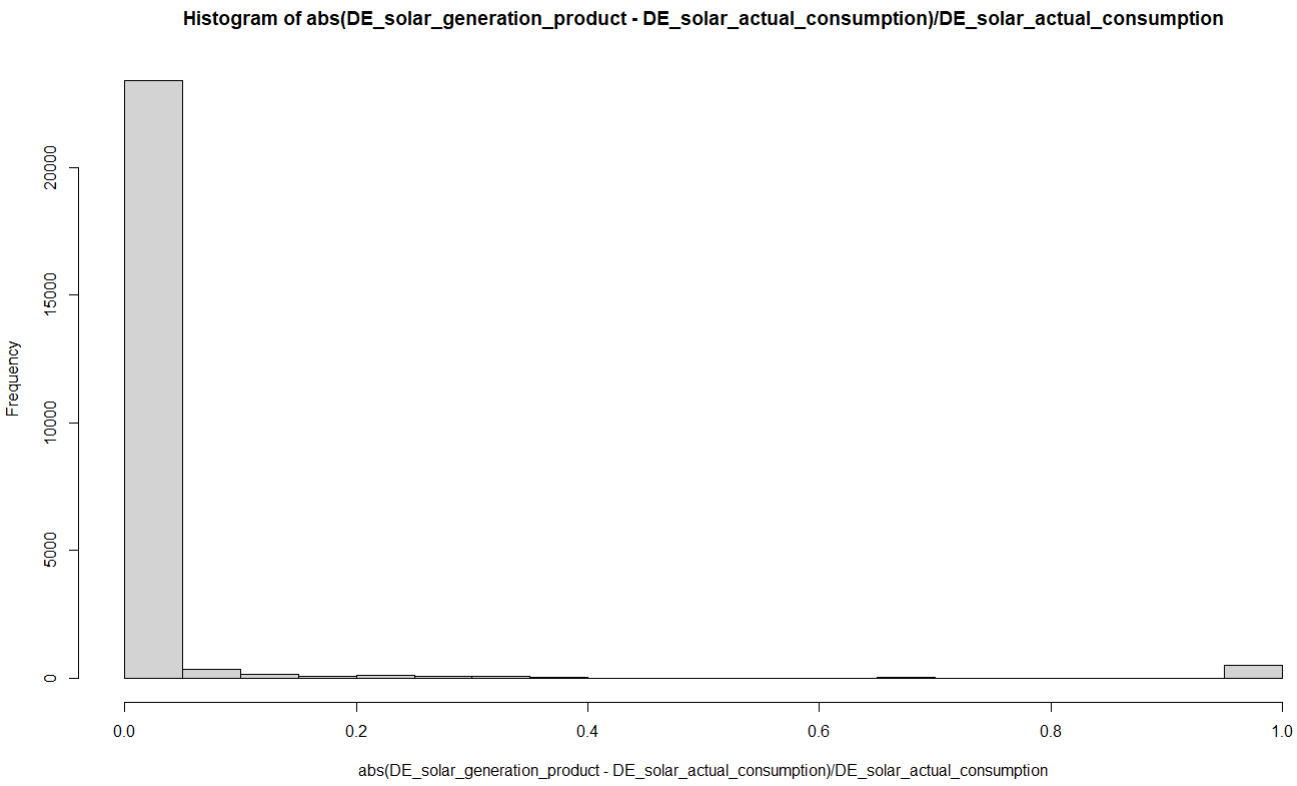

We can actually see here, that for Germany the solar and wind capacities increase over time. So we need to have fixed capacities, we could for example take the average capacities as an arbitrary choice.

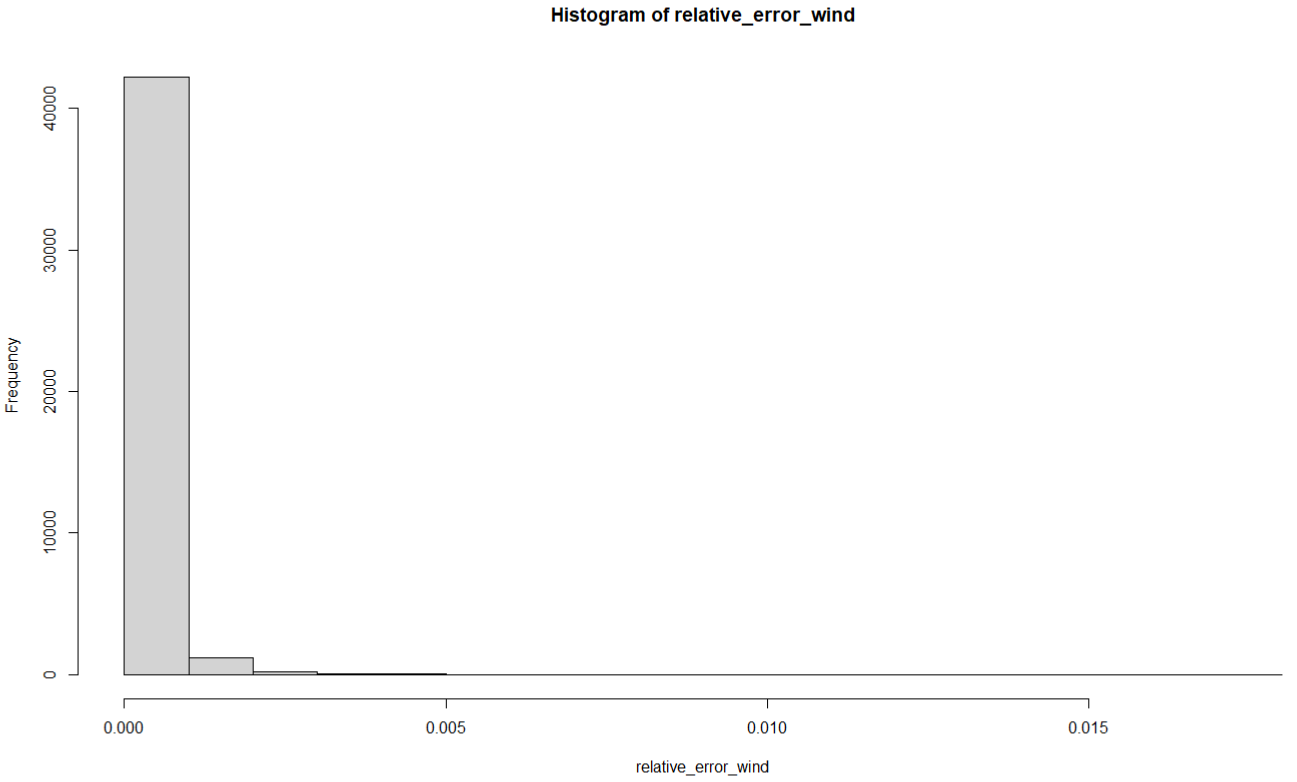
The idea here is to multiply the energy profiles with the fixed capacities, so we will have two columns, one for the available wind energy and the other for the available solar energy.

To see the mistake we are doing by taking the product of profile with capacity instead of the actual generation of energy, we decide to compute the relative difference between the values for each hour, so we actually get a vector of relative mistakes : $\forall i \in [\![1; length(DE\_solar\_profile)]\!]$

$$relative\_error = \frac{|DE\_solar\_actual\_generation[i] - Fixed\_solar\_capacity * DE\_solar\_profile[i]|}{DE\_solar\_actual\_generation[i]} \quad (1)$$

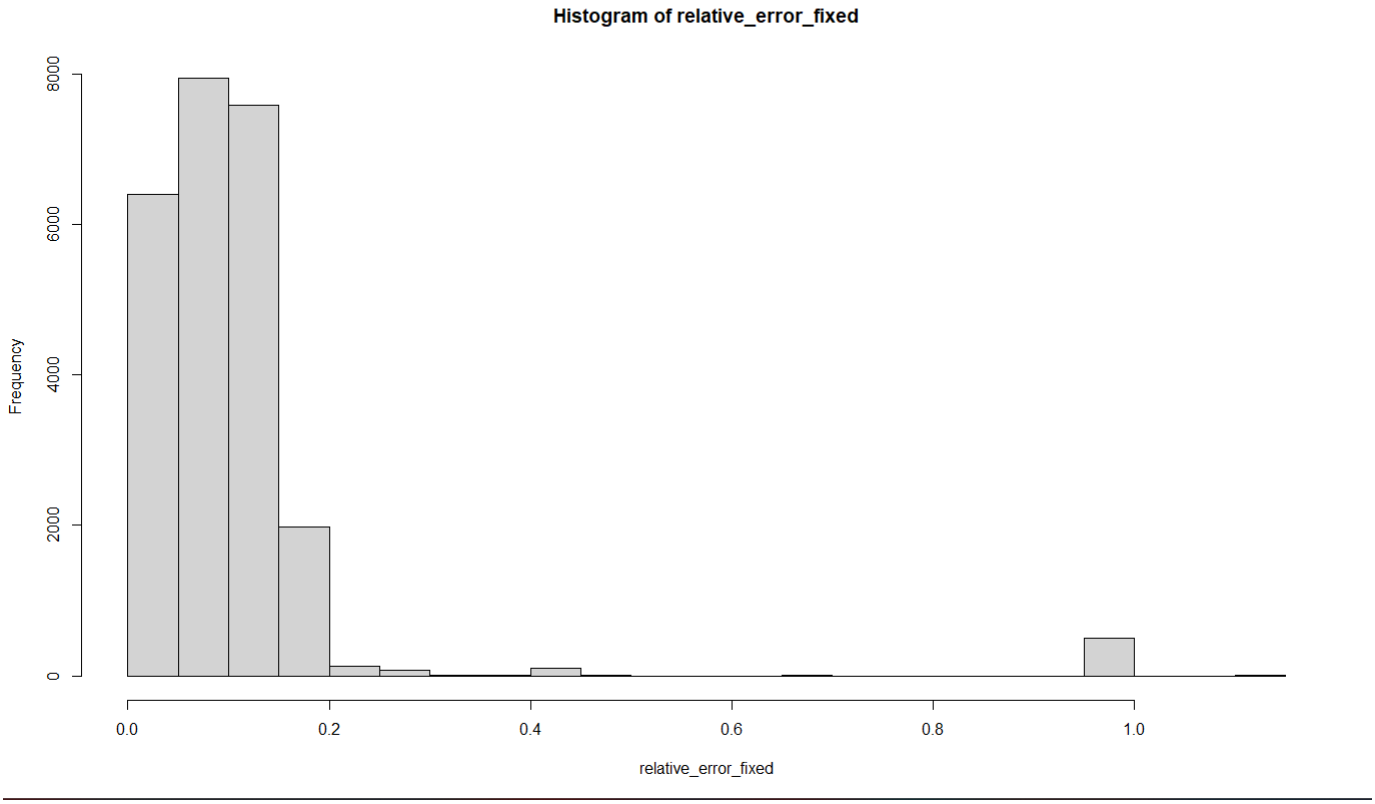
Here we plot the histogram of the relative error, where we took care of the case where the denominator was equal to zero, actually the numerator in these cases was also equal to zero, which in R results in a relative error equal to 1 ($0/0 = 1$ ...).

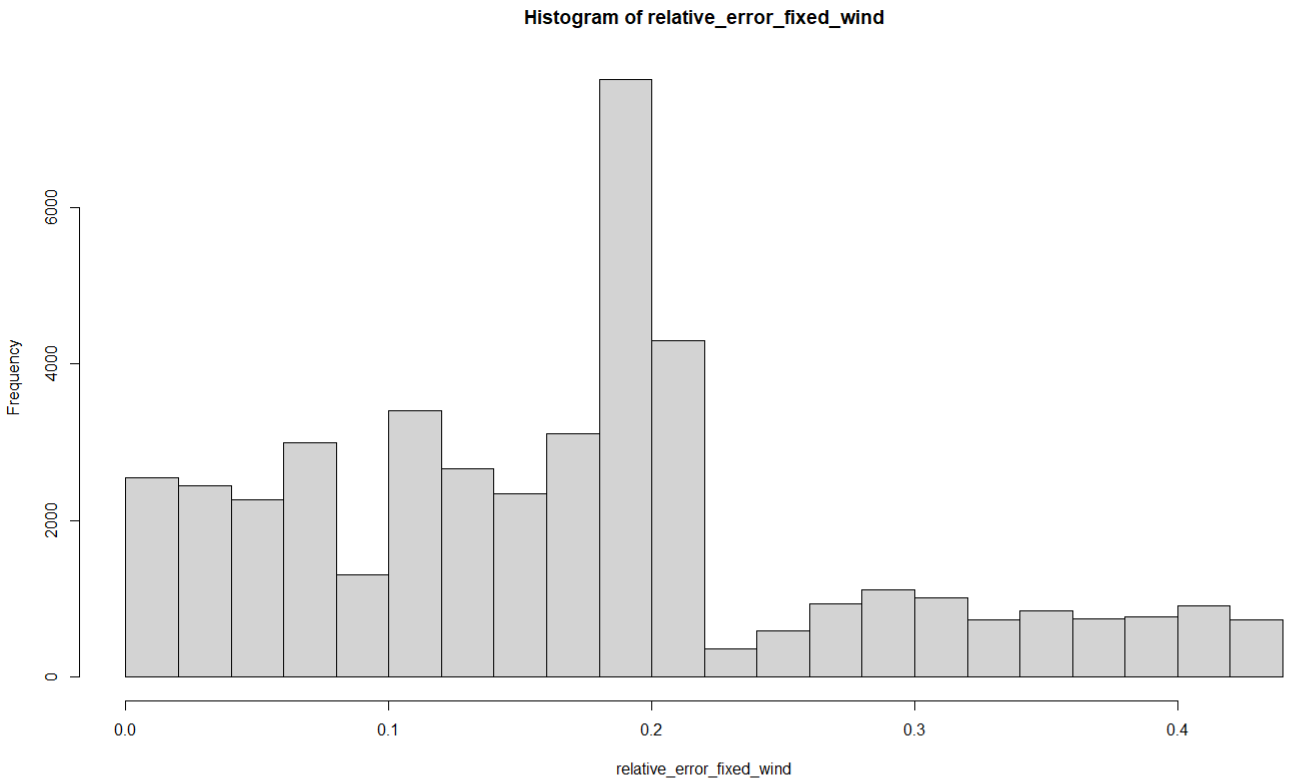

Histogram of abs(DE_solar_generation_product - DE_solar_actual_consumption)/DE_solar_actual_consumption
Frequency
0
5000
10000
15000
20000
0.0
0.2
0.4
0.6
0.8
1.0
abs(DE_solar_generation_product - DE_solar_actual_consumption)/DE_solar_actual_consumption


Histogram of relative_error_wind
Frequency
0
10000
20000
30000
40000
0.000
0.005
0.010
0.015
relative_error_wind

So as we can see, the relative mistake is less than 5% for almost all the components. Which means that replacing the actual energy generation by the product of the profile and the capacity makes sense.

Now, if we take the fixed capacity (here equal to the average of the capacity over time) we get :






So here we can actually see that the relative mistakes are larger than with the actual capacity, but the results are good enough to consider that working with a fixed capacity and these same data makes sense.

Therefore to conclude, we will use the price_day_ahead column of Germany that we will complete with the Austrian one (this is due to the binding zones), in addition we will keep the DE_solar_profile and the DE_wind_profile, which we will multiply with the solar_averaged_capacity and the wind_averaged_cap

## 3.2 Selection of the lines (here it means the dates)

First, we can see that for the four columns we chose : DE_LU_price_day_ahead, AT_price_day_ahead, DE_solar_profile and DE_wind_profile, we still have missing data.

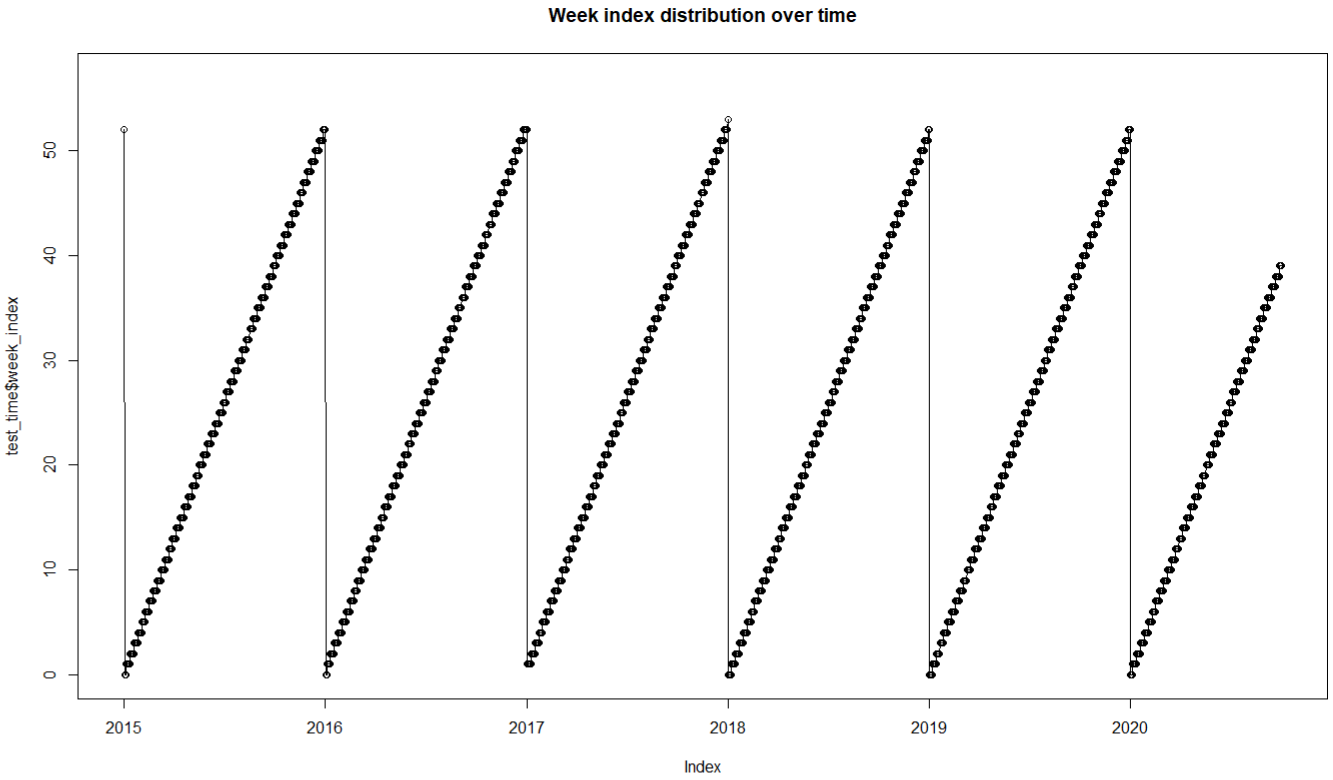
But first, we will work in terms of weeks, that means we will optimize over one week and decide the level of stock of hydrogen we want to have at the end of the week (this will actually correspond to the initial stock of the very first hour of the following week).

So to do this, we add a column which contains the index of the week in the year. In R we do this :

$$df\$week\_index \longleftarrow format(df\$utc\_timestamp, "\%U") \tag{2}$$

"%U": The format string %U represents the week number of the year, where Sunday is considered the first day of the week. It returns a zero-padded decimal number representing the week. All days in a new year preceding the first Sunday are considered to be in week 0. The %U format is based on ISO (8601) week numbering standards.

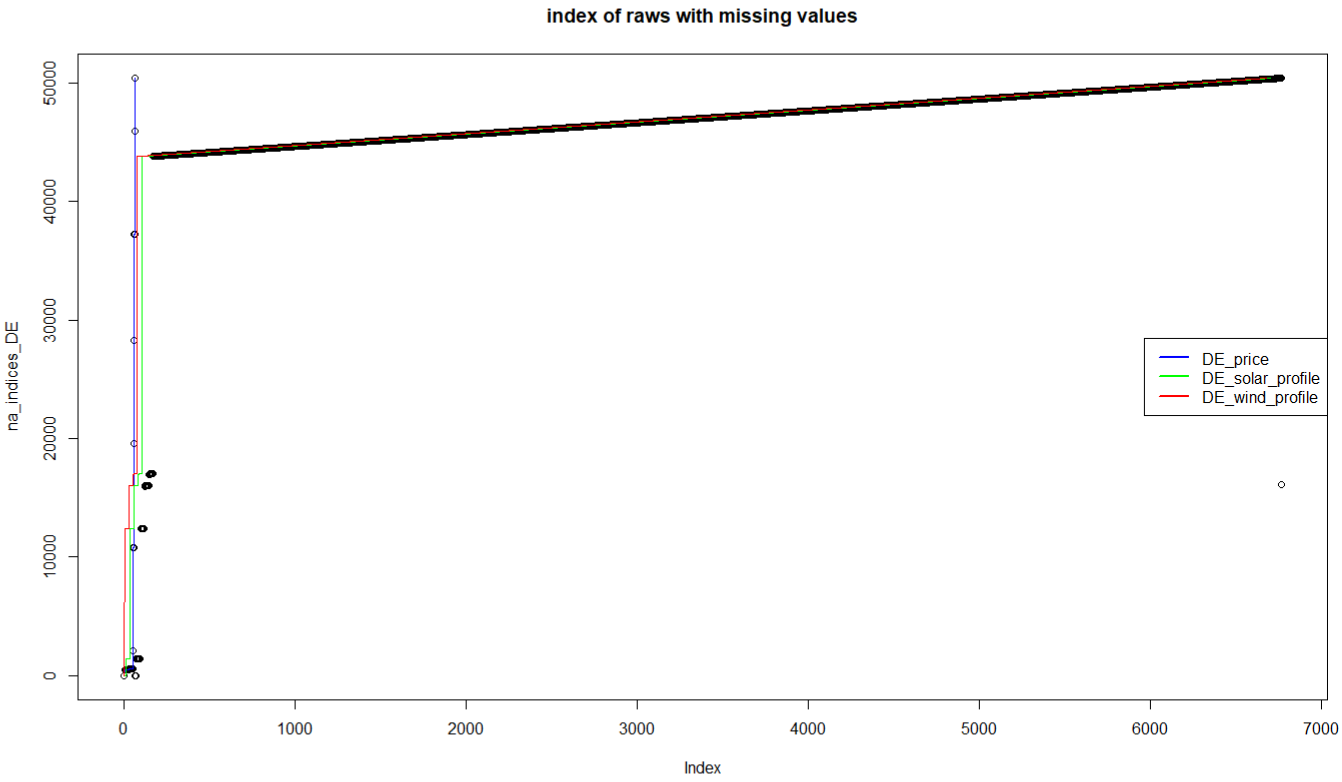
Then we plot this new column to see the distribution of the week index in our database :




This graphic is coherent with our week index transformation.

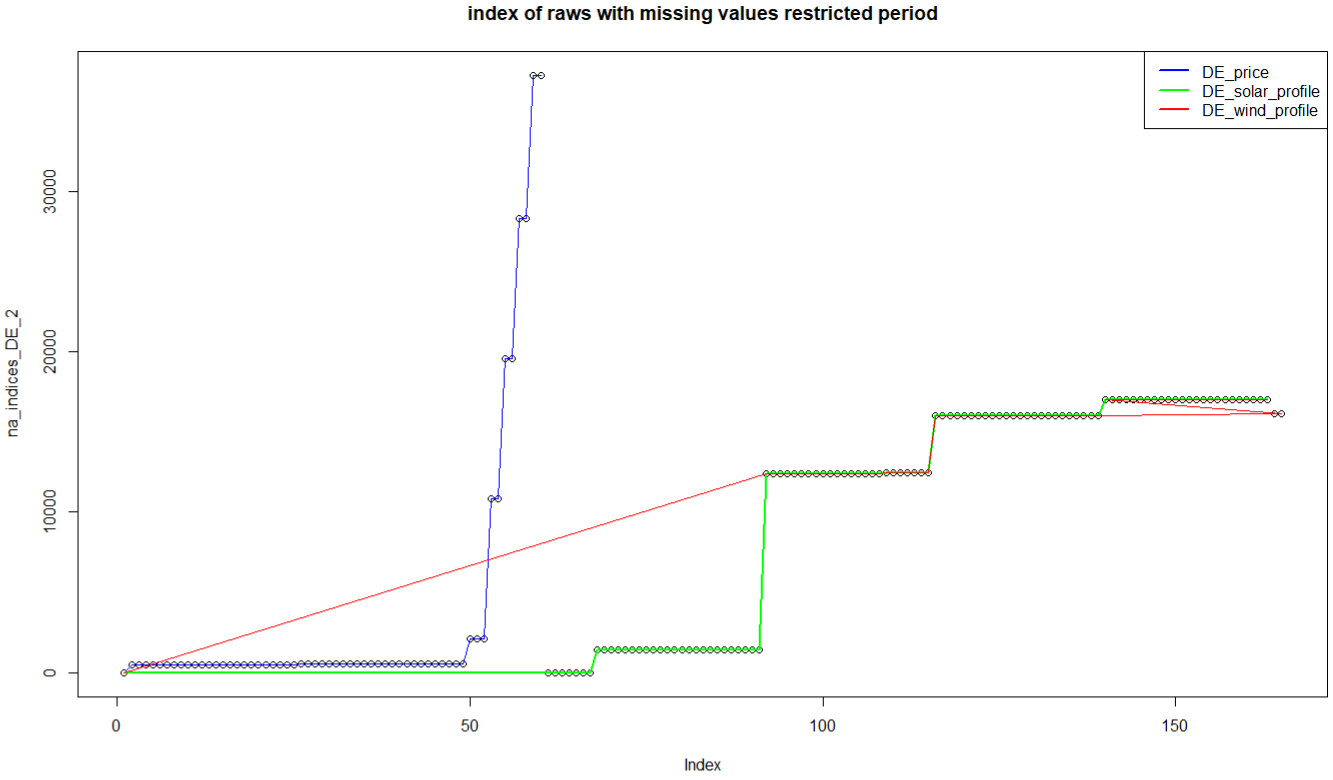
So we have fives years where we can reproduce our experiment. Now we need to select only the weeks which do not contains missing values (noted NA in R).

First, we start with the price, according to the definition of biding zone we need to add the prices of Austria into the Germany column prices where we have missing data, this is the consequence of the separation of the AT_DE_LU biding zone in 2018 into the two biding zones AT and DE_LU.




So here the point which is interesting is that there are only missing value for the DE_solar_profile and the DE_wind_profile around just after the index 43801, which corresponds to 2019-12-30 23:00:00. Consequently, we will restrict on the data before this period.

Here is what we got after restricted the time period:

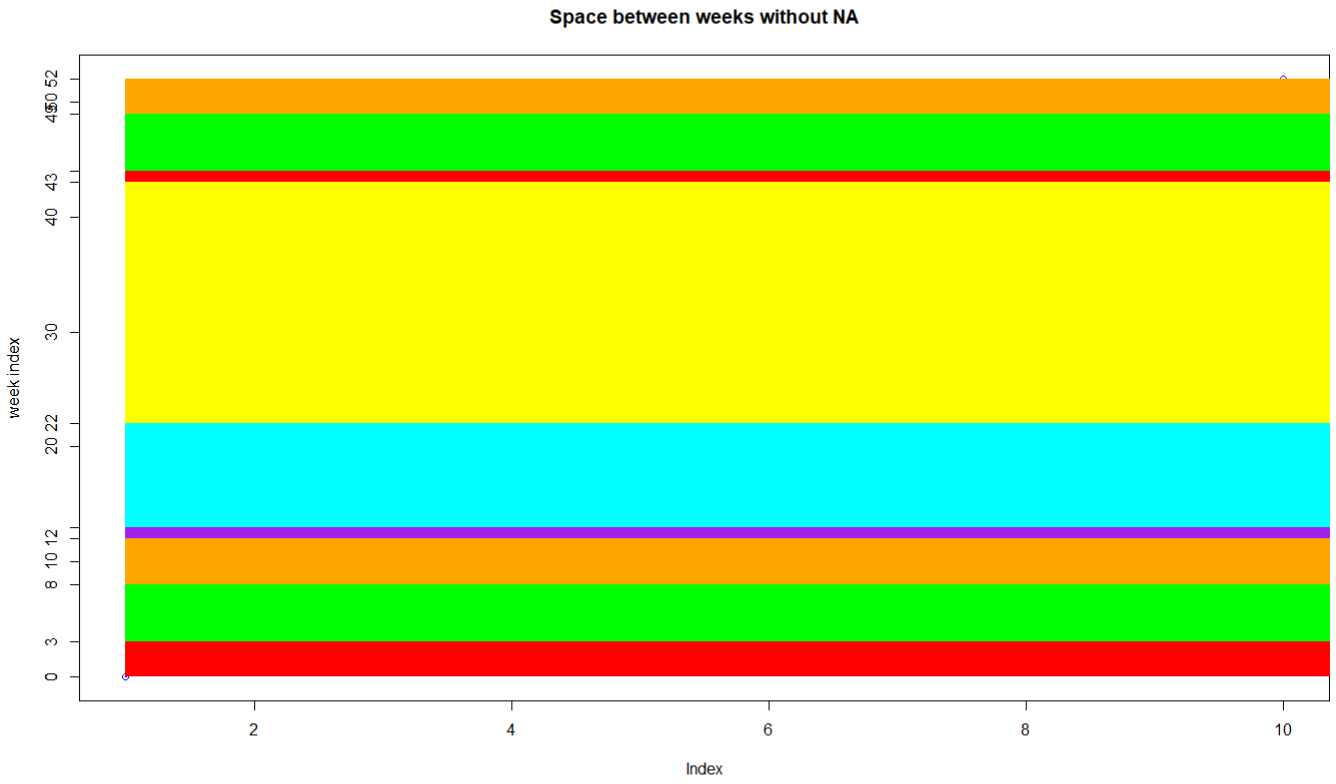



So first we can see that by only restricted the time period we now have 165 rows with missing values instead of 6766 before.

Moreover, we can only see that the missing values are not isolated, indeed except for the price where it seems that groups of to consecutive hours with missing values are present, the other missing values are grouped by around 10.

So here we have different possibilities, first we ignore all the weeks which contain at least one missing value, or we still ignore the weeks with missing values but only in the year where the values were missing, that means we need to be careful while working the indexes of the weeks over the years. A last possibility would be to interpolate the missing values, with for example replacing with the averaged value of the same time (hour-day-month) over the years.

If we chose the first option, here is what we got :




So we have 10 week indexes that contains at least one missing value, this graphic allows to see the number of consecutive weeks we could have without any missing value. This seems actually quite good, regarding the fact that we have here 4 consecutive years, that means we could run our optimization program four times over these same time periods.

To summarize, we would have two periods with two consecutive weeks, one period with 3 consecutive weeks, two periods with 4 weeks, one period with 8 weeks and one period with 20 weeks.

Now if we decide to interpolate (we kind of consider that the missing value here are due to a problem during the recording and not due to a problem with the installation, otherwise we should replace the missing value by 0, but since these data are for all Germany, the recording problem seems definitely more likely):

To do so, we first add a column which is the same as the utc_timestamp column but without the year, this will allow to interpolate over the years. Then we use the which() function in R which allows to create a list of integers which correspond to the index of the vector which fill the condition, in other words here we retrieve for example the wind values which correspond to the same date of the missing value but at a different year:

```
df_test$MonthDayHour <- format(df_test$utc_timestamp, format = "%m-%dT%H:%M:%S", tz = "UTC")

date_wind_profile_NA <- df_test$MonthDayHour[na_indices_DE_wind_profile_2]
DE_wind_profile <- df_test$DE_wind_profile[1:43800]
for (i in 1:length(date_wind_profile_NA)){
  date <- date_wind_profile_NA[i]
```

```
  index_date <- which(df_test$MonthDayHour == date)
  wind_value <- DE_wind_profile[index_date]
  test_time$DE_wind_profile[na_indices_DE_wind_profile_2[i]] <- mean(wind_value,na.rm = TRUE)
}
```

After checking, we can conclude that these columns of price, solar and wind profiles do not have any missing value anymore.

So we will create a new database where we actually keep the two options.

Here is a picture of the finale database for Germany :

SQL ▾ < 1 / 876 > 1 - 50 of 43800

| utc_timestamp | DE_week_index | DE_price | DE_solar_profile | DE_wind_profile | DE_price_with_NA | DE_solar_profile_with_NA | DE_wind_profile_with_NA |
|---|---|---|---|---|---|---|---|
| 2014-12-31T23:00:00Z | 52 | 19.64 | 0 | 0.471925 | NULL | NULL | NULL |
| 2015-01-01T00:00:00Z | 0 | 35 | 0 | 0.3171 | 35 | NULL | 0.3171 |
| 2015-01-01T01:00:00Z | 0 | 45 | 0 | 0.3244 | 45 | NULL | 0.3244 |
| 2015-01-01T02:00:00Z | 0 | 41 | 0 | 0.3249 | 41 | NULL | 0.3249 |
| 2015-01-01T03:00:00Z | 0 | 38 | 0 | 0.3283 | 38 | NULL | 0.3283 |
| 2015-01-01T04:00:00Z | 0 | 35 | 0 | 0.3307 | 35 | NULL | 0.3307 |
| 2015-01-01T05:00:00Z | 0 | 35 | 0 | 0.3471 | 35 | NULL | 0.3471 |
| 2015-01-01T06:00:00Z | 0 | 36 | 0 | 0.3701 | 36 | NULL | 0.3701 |

# 4 Study of the impact of the final constrained level of stock

## 4.1 Study of the methods

We consider that we are working in a non deterministic context due to the randomness of the renewable energy through the solar panels and the wind turbines. So we want to produce hydrogen using almost only renewable energy. That means we have a demand that we have to fulfill, to do so we produce hydrogen or we use the quantity we produced before and stored. But to produce hydrogen we need electricity, so if at some time we do not have enough stored hydrogen and enough electricity from the PPA to fulfill the demand, we will have to use electricity from the grid, which is expensive. But if we had stored enough hydrogen we would not have had to buy electricity from the grid, so we have not been good enough at predicting our renewable energy data, thus we did not anticipate the big storm that happened and we did not stored enough hydrogen.

It is actually here that we can have two different ways to address this issue. The first one would be to start a week with all the information known in advance, that means we begin our week with our plan for the production, that means we already know exactly what the production of renewable energy will be, if a storm should have happened we will already have taken into account this event in our plan. Then we just constrained the final level of hydrogen available at the end of the last hour of the last week, which will actually correspond to the initial level of the next week. Then we begin our new week with a certain level of hydrogen stored in a tank, we again know all the information of this new week, so we create our optimal production plan taking into account the initial stock of hydrogen.

Here, we consider that we only know the information of a week and we decide to have a certain level of hydrogen stored in a tank at the end of the week, we have therefore no idea of what could be the following week hence the importance of the final constrained level of hydrogen. The problem of this context is that it is not

very realistic, indeed, although we do no know exactly what will be the production of renewable energy of the following week, we could still use good prediction. This leads to the second method :

We could do our production with one deterministic week as a first week, and then we would take two random weeks which are likely regarding the first week chosen, and then we would construct our optimal production plan. The idea then would be to compare the optimal cost obtained by optimizing over the three weeks (one deterministic an the two other predicted by a model or chosen randomly) to the optimal cost obtained over the three true weeks. So we still constrain the level of hydrogen at the end of each week but, the result will be more realistic.

With the second method, we fix the number of weeks Nw over which we build our optimal production plan : we choose a first true week, then we predict Nw-1 weeks, we also store the true Nw-1 weeks. Then we run our optimization program in the two cases, where for the first case we constrained the final level of hydrogen with a method to define later, and we compare the two optimal costs obtained by running this optimization program.

So first we need to choose the first true initial week. To make things easier to start we will consider only time period over a year, that mean we cannot have the first initial week in December and considering more than 4 weeks, otherwise this could be a problem with the indexation, but it could be resolved later.
So after fixing the time period, for example we want to start at the 37th week work until the 43th week. We need to choose the true year, the one from which only the first week will be extracted, then we need to complete the 5 following weeks, to so there are different possibilities, either we used the 38th, 39th,..., 43th weeks from other years picked randomly, or we can use a model that will have the first week as an input and will be able to predict the 5 following weeks.

## 4.2 The problem of negative prices

Source : https://www.cleanenergywire.org/factsheets/why-power-prices-turn-negative

"Negative power prices on the electricity exchange occur when a high and inflexible power generation appears simultaneously with low electricity demand. This is often the case on public holidays such as Christmas or Pentecost. Particularly in hours of (predictable) high renewable power supply (lots of wind and sun), power producers offer their electricity for negative prices on the exchange. This is often done by marketers of renewable power but also by conventional power stations like nuclear and lignite plants. In this event, the market clearing price can be set below zero."

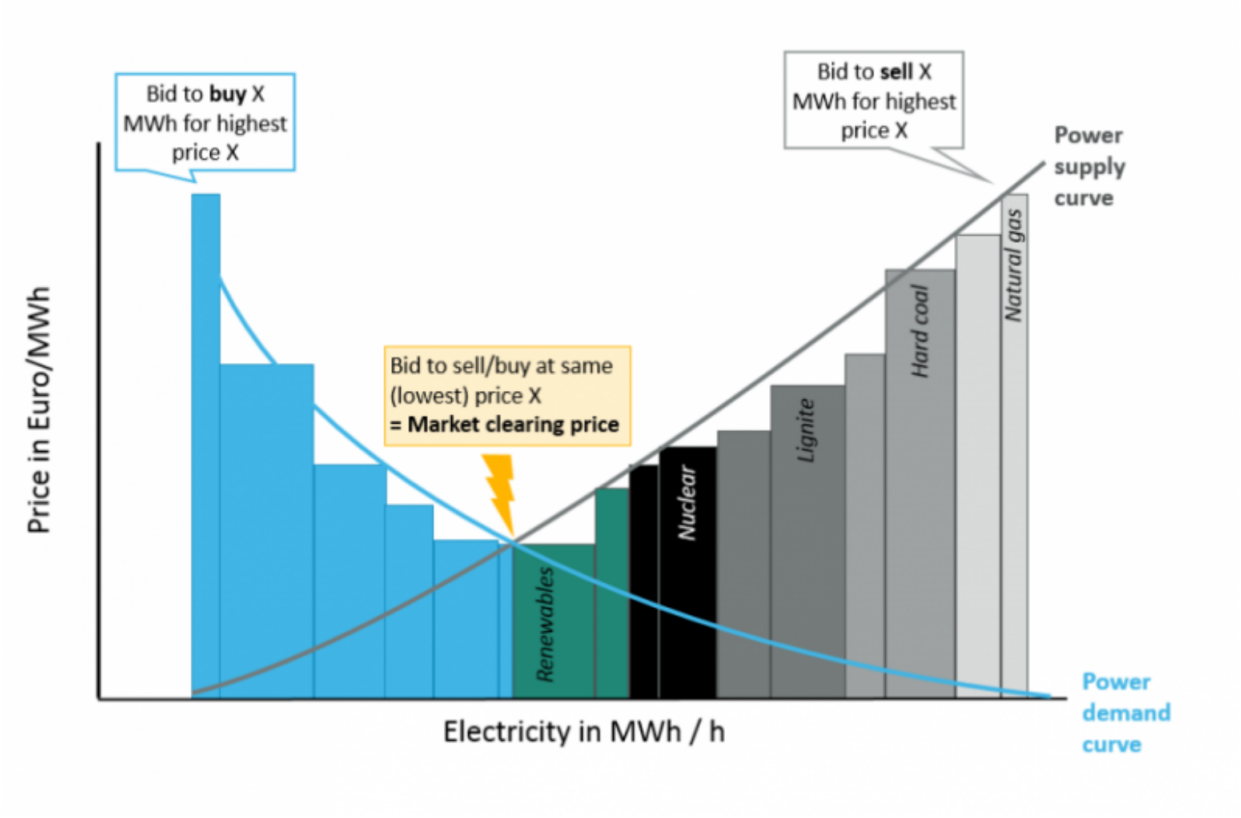



Figure 1| Diagram of how the power price is found at the electricity exchange. Power station operators bid their respective amounts of power for every hour of the day into the market at a certain asking price. Buyers lodge their bids. The merit order leads to the cheapest offer determining the market clearing price.

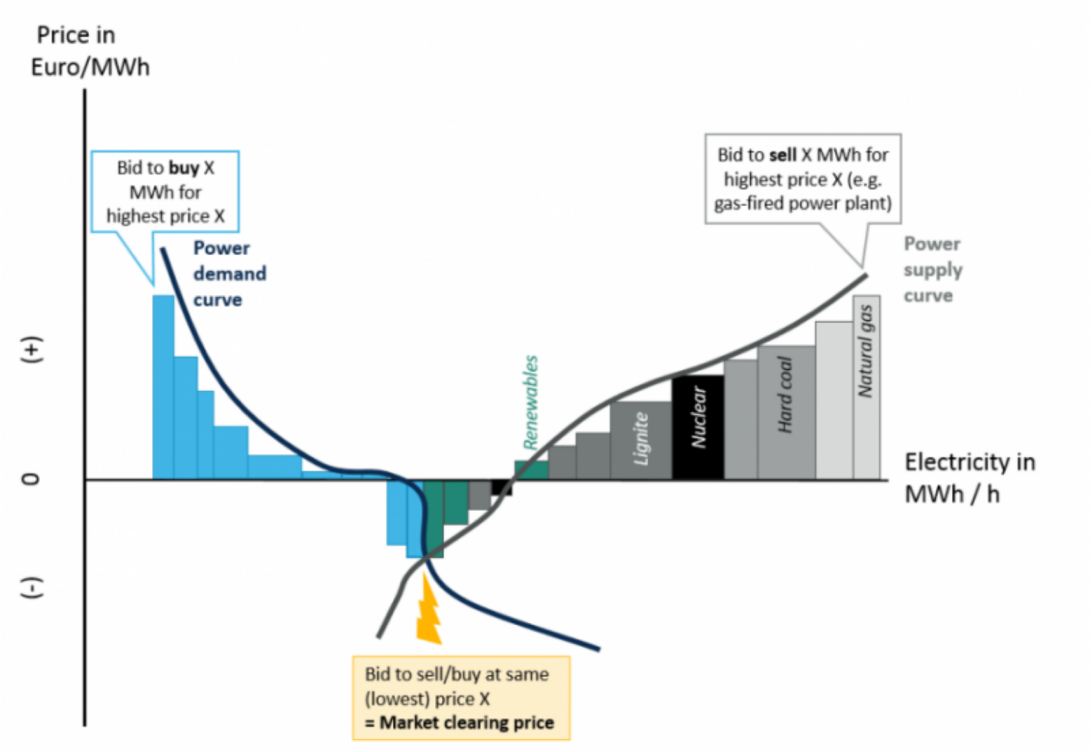
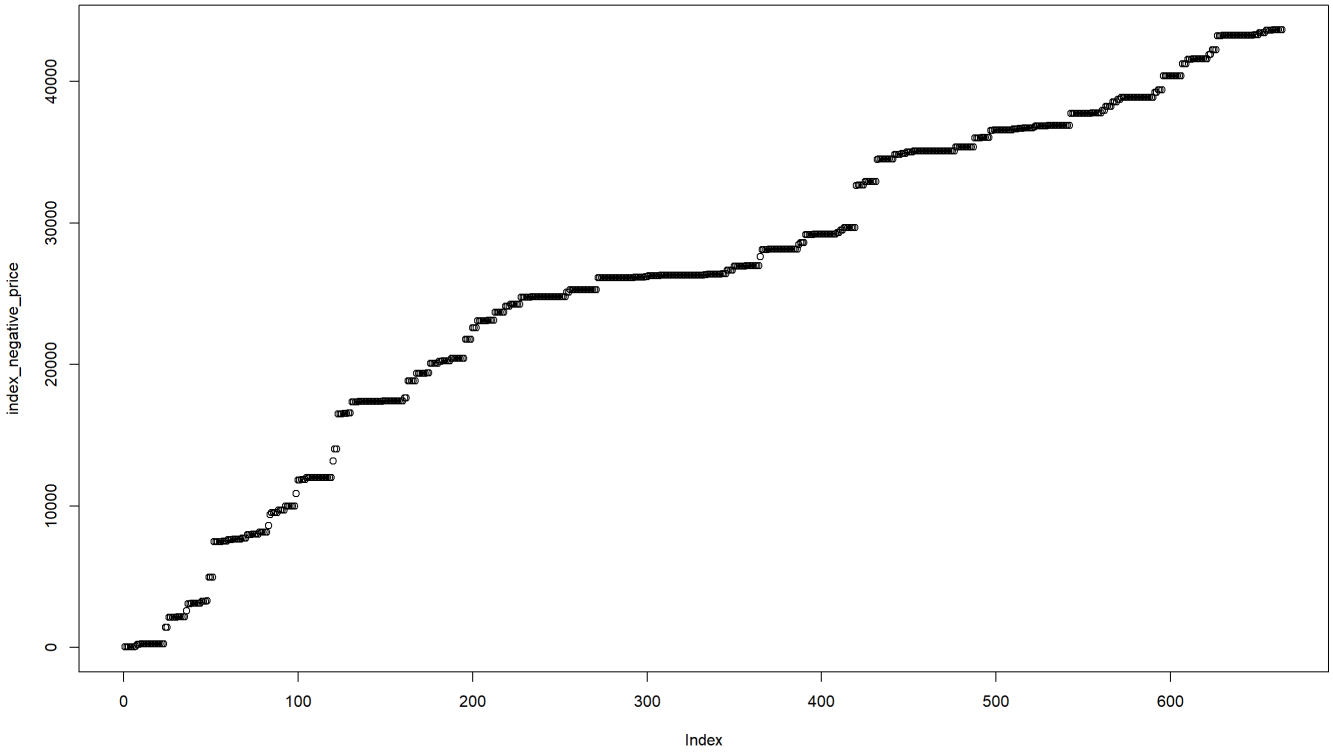



Figure 2| Diagram of power market situation that leads to negative market clearing price.

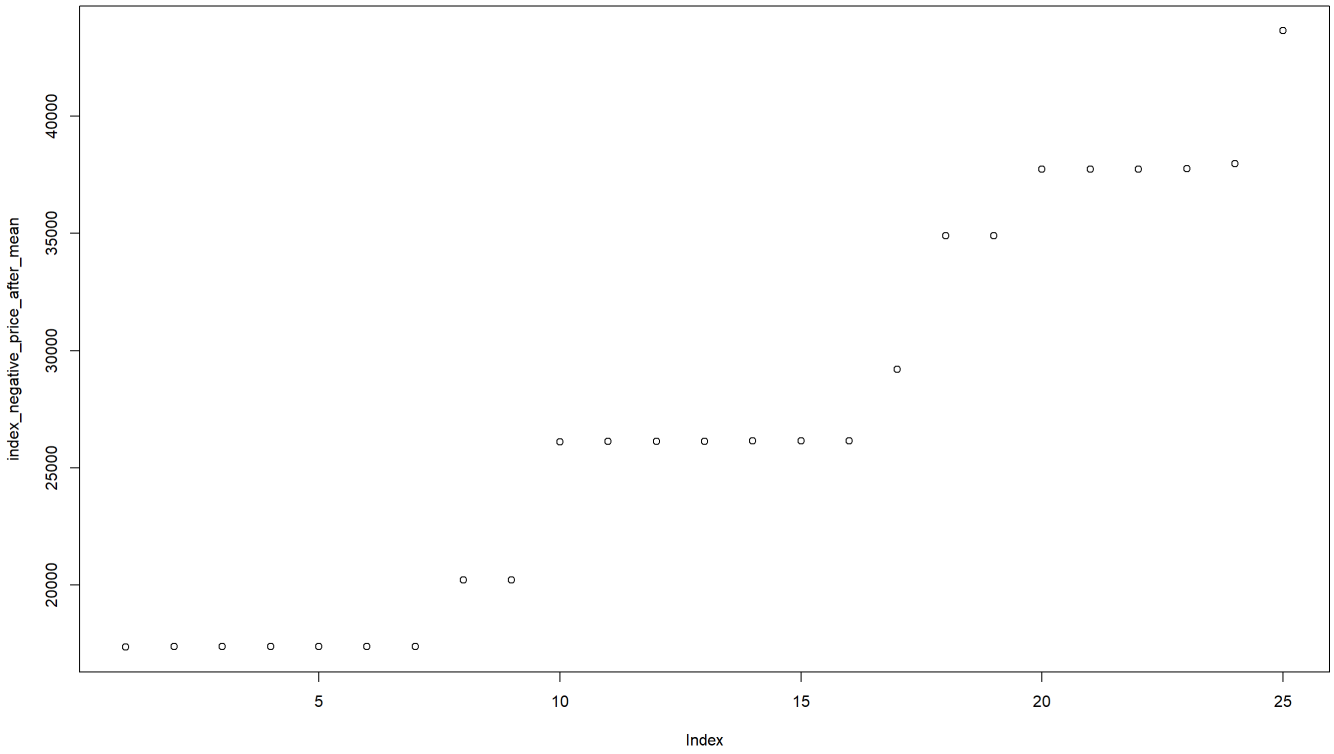
First it seems not realistic to predict negative for electricity in a production plan or maybe it is not that bad. The fact is it is said that a predictable high renewable power supply can happen, but using this sporadic events in prediction for a production plan seems a bit unrealistic, but not that absurd, since even if the price is not negative it could be very low, and so in normal cases the optimal production plan would have anyway prioritized these low-price hours.
But if we use prediction models, then these sporadic events should disappear.

Here is what we have with the German prices (Germany is impacted a lot by these negative prices, since it has a lot of renewable energy) :



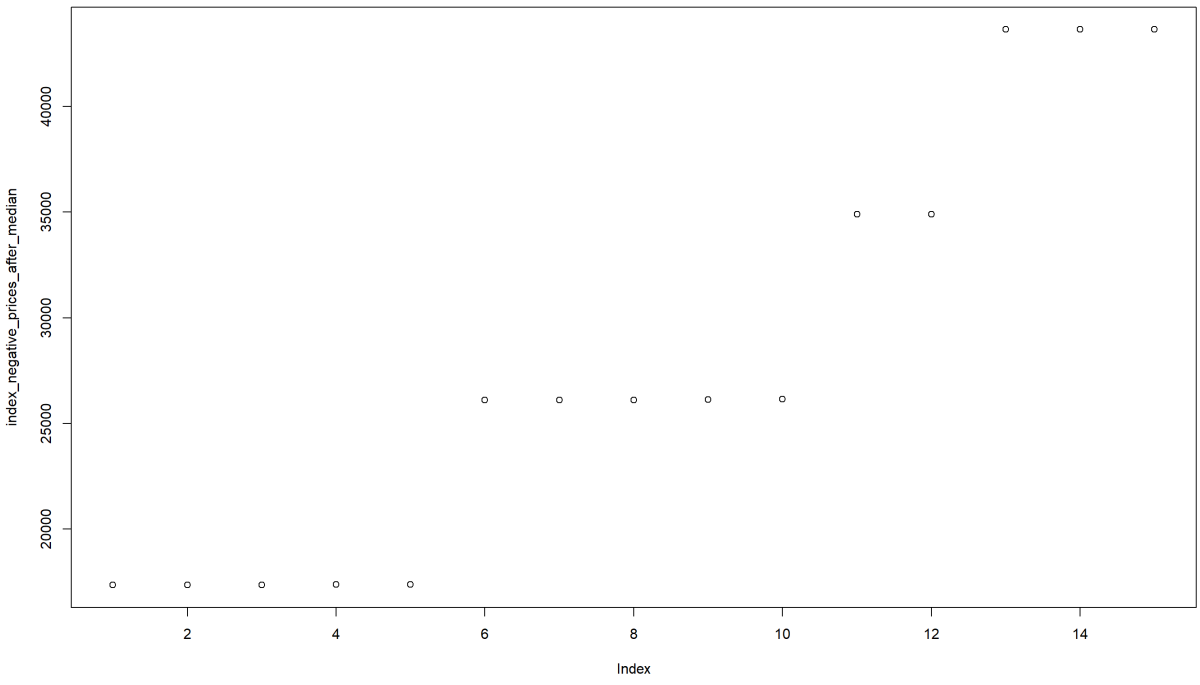

So at this time, we add a column where the negative prices are replaced by the average value over the years as we did previously.

Here is what we got :




What is very interesting is that we still have negative prices after averaging, this would mean that the negative prices are not as sporadic as it could have seemed. Or it could be that at some hours the prices were very negative, and consequently the average did not get rid of this problem. So instead of using the mean, we could actually use the median.

Here is what we got :



What is very interesting is that we still have negative prices but less. We can also see that we do not have isolated point anymore.

So we will actually keep the negative prices after taking the median, because that means they could have some statistically relevance, and having 15 negatives prices over 43800 prices is not that significant.

So to be clear, in the construction of the optimal production plan with the unreal weeks, we will use the prices after taking the median, but for the computation of the optimal cost in the real case we will use the real prices.

# 5 The problem of transposition of the production plans

## 5.1 Explanation of the problem

Actually we have a problem of transposition between our true situation and our predicted situation. Indeed, the comparison of the two optimal costs does not really make sense, since we are in two different situations. The idea would be to use our predicted weeks to get a production plan where we actually constrained some final levels of hydrogen, then we transpose this production plan into the real situation.

The problem is that our predicted production plan may not be feasible in the real situation, for example if we planned to used 70MW from the PPA at 12 o'clock on a Saturday, but there are only 40MW which are available, then we need to review our production plan, but the question is how much we have to review this plan, could we just not only use 30MW more at the following hour? But if it is not possible, can we do it at the second following hour? But if we had been so bad at predicting and this consequently resulted in being unable to fulfill the demand, then we have a problem. We could imagine tremendous cost to punish the incapacity of fulfill the demand instead of taking it as a constraint. So the point is that we need to create a transposition model, with costs for changes and not fulfilling the demand, these would be mistake-costs.

So the idea would be to create a second optimization problem where we try to make our predicted production plan feasible with as low mistake-costs as possible. Therefore we first need to define the possible changes we can make, for example can we at 1pm on a Saturday use 50MW instead of 20MW as we decided in our predicted production plan, and then should we actually consider a cost for this change and if so what should be the cost?

There is also the problem of the momentariness of the mistake. That means, we built our production plan with predicted data for energy, we begin our working period, at the beginning everything is fine, the production plan may not be optimal with the true data but it is still feasible (this case could be for example more likely, if we used upper bound for the energy generated by the PPA, for instance we predict 70 MW of solar energy available at 12 o'clock on a Saturday, but we constraint the model to use at most 50 MW, we could use some standard error to create good upper-bound), but suddenly the production plan is not feasible anymore (for example not enough energy available from the PPA). We could say that this is not actually a problem, as long as our production plan is feasible we do no modify it, but since actually the in-feasibility can only come from a lack of energy, we can for example use all the electricity available and move the consequences to the following hours. But we need to deal with the mistakes in a chronological order.

Furthermore, we are back to one of our previous problems namely the time when we can build our optimal production plan. Indeed we do not have enough available energy at 12 o'clock so we decide without any thought to use more energy at the next hour to dealt with the lack of energy. But for example we can actually imagine that the energy at the following hour was very expensive, so a better idea would have been to wait until the second following hour to use more energy if it was possible. Nevertheless, that means we are able to optimize

our choices in the middle of a week, so according to the optimization power we would have at that time, it could even be better to just create a new optimal production plan, and then we are actually back to the problem presented at the beginning of this paragraph.

Consequently, to keep some coherence with our model, the best choice would be to take the easiest choice, which we still need to define, but this will probably amount to changing the consumption of energy from for the PPA at the following hour, or are the closest hour when we can actually use more energy.

Actually, we will just use the energy from the grid, which is definitely an easier way to deal with these false predictions. Indeed, we suppose here that the grid has infinite capacities, the only problem is that its electricity is costlier than the one from the PPA. So if we predicted and used more electricity from the PPA that was actually available, we complete with electricity from the grid which will increase our final cost.

## 5.2 The function of the transposition of the cost

Here we develop how this function of transposition is designed.

As explained in the previous part, a production plan is computed with an optimization program based on predicted data. So the cost given by this program does not correspond to the real one, that is the one that will be obtained by applying the given production plan with the real prices on the market and the real quantities of energy available from the PPA.

Furthermore, as decided previously to transpose our production plan we have only the possibility to manage the origin of the used electricity at a hour. If we predicted more electricity from the PPA that is actually available we complete with electricity from the market. But if in our predicted production want electricity from the market is used at a certain hour although in the real case there is still electricity from the PPA that can be used, then instead of using electricity from the grid we use in priority the one from the PPA and after it is not enough we complete with the grid.

This two considerations lead to the actual consumption of electricity with our production plan in the real case. Here is therefore finally the function of transposition of the cost which gives the real cost, and which will be the objective function noted F that will be as minimized as possible through the different final levels of stored hydrogen :

$\mathbf{N_w} \in \mathbb{N}^* \setminus \{1\}$ : number of weeks over which the predicted production plan is computed

$\mathbf{x_s} \in \mathbb{N}^{N_w - 1}$ : vector of the constrained final levels of stored hydrogen.

$\mathbf{Opt\_PPA}(x_s) \in \mathbb{R}_+^T$ : represents the quantity of electricity from the **PPA** used at each hour according to the predicted production plan which depends on $x_s$.

$\mathbf{Opt\_Ener\_Market}(x_s) \in \mathbb{R}_+^T$ : represents the quantity of electricity from the **grid** used at each hour according to the predicted production plan which depends on $x_s$.

$\mathbf{Real\_PPA} \in \mathbb{R}_+^T$ : represents the quantity of electricity generated by the PPA at each hour in the real case.

$\mathbf{Real\_price} \in \mathbb{R}^T$ : represents the prices on the grid at each hour in the real case, here prices can be negative as discussed previously.

$$
\begin{aligned}
F(\text{Real\_PPA, Real\_price,Opt\_PPA}(x_s)\text{,Opt\_Ener\_Market}(x_s)) &= \sum_{t=1}^{T} [\text{Opt\_PPA}(x_s)_t - \text{Real\_PPA}(x_s)_t]^+ \text{Real\_Price}_t \\
&+ \sum_{t=1}^{T} [\text{Opt\_PPA}(x_s)_t - [\text{Real\_PPA}_t - \text{Opt\_Ener\_Market}(x_s)_t]^+]^+ \text{Real\_Price}_t \\
&= \sum_{t=1}^{T} [\text{Opt\_PPA}(x_s)_t + \text{Opt\_Ener\_Market}(x_s)_t - \text{Real\_PPA}_t]^+ \text{Real\_Price}_t
\end{aligned} \tag{3}
$$

# 6 Study of the impact of the final levels of stored hydrogen

## 6.1 First considerations

Here we present the algorithm with the different functions which will be used to study the impact of the final constrained level of hydrogen.

The first concern here is about the costs of curtailing, these costs can have a significant impact on the final cost, this depends a lot on their modeling. Indeed, let us assume we are in the prediction part, at an certain hour on a certain day we predict a lot of electricity available from the PPA, we run our optimization program based on these predicted data, and it actually says that the optimal predicted production plan does not need to use whole the electricity from the PPA at this certain hour. So we have a surplus of electricity from the PPA at this certain hour, it is known that producing more electricity that is consumed is costly, hence the negative prices, but this surplus of electricity could also be sold on the market. So either we penalize this surplus of electricity from the PPA not used by the predicted production plan, in order to actually force the optimization program to avoid unused electricity from the grid, or we take benefit from this surplus of energy and sell it on the market, which is in fact the opposite way to deal with this unused energy. the only issue with this second method is that it supposed that the electricity from the PPA can actually be sold on the market, which can seem as a non-negligible hypothesis.

The first method has its drawbacks. Indeed, let us assume we penalize the surplus of electricity from the PPA, then the program of optimization will give a predicted production plan which used as mush as possible the electricity from the PPA, which can lead us to a very far optimal predicted cost compared to the real case except if we use upper bound for the available electricity from the PPA as it was previously said. To my opinion the biggest issue is more that this method could make the analysis too much far from the second method. Indeed in that case, it could be actually a good idea not to use when the electricity from the PPA and sell it on the market when the prices are high, and with the first method this concept does not appear.

To summarize, I would say both methods could be say to be relevant, so a variant of the first method was chosen. The idea is to consider that the electricity from the PPA is free compared to the one from the grid whose price is hourly dependent. This hypothesis could be realistic if the solar and wind farm are part of the whole infrastructure used to produce hydrogen. This is not exactly the same as the first method, it is a bit different. In fact, let us assume we know the demand for the whole period, so we know the quantity of electricity we need (actually because of the phenomenon of dissipation of the battery, this is not really true, since if our program of optimization stores electricity in the battery, it is not the same thing as just directly use t to produce hydrogen, because a part of this stored electricity will have disappeared by the next hour), so let us rather say we have a lower bound for the electricity we need to fulfill the demand. Let us imagine this lower bound turns to be lower than the total quantity of energy generated by the PPA, so the production plan is forced to use electricity from the grid. The fact is if the surplus of electricity from the PPA is penalized another production plan could be given by the program. Indeed let us imagine the case where we want to use all the available electricity from the

PPA which has as a consequence the saturation of all our storage capacities : the battery and the tank, so at the next hour when the electricity from the grid was very cheap, an idea would have been to use this electricity to store more hydrogen or electricity in the battery, than having to use electricity from the grid at another time when the price were actually very high, which will at any case have to be done, since the hypothesis was that the total energy from the PPA was not sufficient to fulfill the demand without relying on the grid. So the best choice would have been to not use all the electricity from the PPA at the previous hour in order to be able to use electricity from the grid at this hour.

This case could seem a bit too exaggerated, but with previous simulations of production plan, it has been observed that the capacities do tend to be saturated at some time during the period. But this is also differ,t from the second method since here the electricity from the PPA cannot be sold. Nevertheless the problem is that this case is not realistic, it is actually the same thing for the first method, since there could be a case where the total electricity generated by the PPA is more than the quantity needed to fulfill the demand, then there will be penalization for this unused electricity, but this electricity will still be there. SO the interest of the second method is that it is the most realistic one, since the unused electricity does not disappear, it is instead sold on the market.

We also decided not to take into account the costs of hourly changes in the production. The main reason here is that we want to keep the simplest model as possible. But having a smoother production could make sense depending on the infrastructure, we could for example imagine a case where the electrolyser will be complicated to pilot within two hours, then to change the production we could have to stop it for a few minutes an then to restart it, which could actually have a cost.

Moreover, there is also a issue for the last week of the period we want to optimize on. Indeed, in our modelling we have Nw weeks, then we decide to constrain the final level of hydrogen of the first Nw-1 weeks and let the final level of stored hydrogen of the last week be. This is more like a piece-wise continuous optimization over time, after Nw weeks we stop and then we decide to create whole new predicted production plan for the following new NW weeks. This does not seem to be a major issue now, this is why the last weeks is unconstrained for its final level of stored hydrogen, in a certain way we always assume, that this unconstrained level will be null, since it will not be optimal to have produced more hydrogen that was actually needed (this hypothesis is reinforced by the fact that we do not penalize the curtailing of the electricity generated by the PPA, otherwise this could led to produce more hydrogen that is actually needed since there are no OPEX).

After these considerations, we can now present our algorithms to study the impact of the constraints of the final stored levels of hydrogen at the end of the Nw first week of a fixed period.

## 6.2 Pseudo code for the algorithms

Our study is split into three parts :

1) The generation of predicted data: we fix the time period and then we generate prediction for the available energy from the PPA and for the prices on the grid.

2) The creation of the optimization problem with the predictions: this is here that all the modelling takes part with the constraints and the chosen objective functions which is composed of different costs (here we only consider the costs of energy from the grid as explained before).

3) The use of a heuristic : this is where the function of the transposition of the cost takes part, the aim will be to find the lowest real cost that could be obtained with different values for the constraints of the final levels of stored hydrogen.

### The generation of predicted data

As explained before the data used for this study are for Germany from January 2015 to December 2019, so there are 5 complete years. The indexes of the weeks goes from 0 to 52 (except in for 2018 where it goes to 53, so we will keep indexes up to 52 for our predictions). We will for our first study use the method of prediction explained before : the idea is to use a true week then to predict the four following ones by picking them randomly among the years. There are some drawbacks for this method, first we assume that we can perfectly predict the first week for our optimal production plan, which is definitely unrealistic. Secondly, the method of prediction itself is not also really realistic, it should be rather seen as a default way of predicting data.

We will denote :

$\mathbf{N_w} \in [\![2; 52]\!]$ : here we restrict to the range $[\![2; 52]\!]$ because of the choice of our model of prediction

$\mathbf{N_h} = 168N_w$ : total number of hours of the period

$\mathbf{FP} \in [\![2015; 2019]\!] \times [\![0; 52 - N_w]\!]$ : the fix period (the year and the week index)

$\mathbf{DE\_solar\_profile} \in \mathbb{R}_+^{\text{len(DE)}}$ : represents the part of solar energy used for the whole consumption of electricity per hour.

$\mathbf{DE\_wind\_profile} \in \mathbb{R}_+^{\text{len(DE)}}$ : represents the part of wind energy used for the whole consumption of electricity per hour.

$\mathbf{DE\_Price} \in \mathbb{R}_+^{\text{len(DE)}}$ : represents the price of electricity on the German market per hour.

$\mathbf{DE\_Year\_index} \in \mathbb{R}_+^{\text{len(DE)}}$ : contains the year of each row of the dataset (see 3.2.Selection of the lines).

$\mathbf{DE\_Week\_index} \in \mathbb{R}_+^{\text{len(DE)}}$ : contains the index of the week of each row of the dataset (also see 3.2.Selection of the lines).

$\mathbf{Solar\_capacity} \in \mathbb{N}$ : represents the maximum of electricity that can generate the solar farm of our infrastructure (see 2.Size calculation).

$\mathbf{Wind\_capacity} \in \mathbb{N}$ : represents the maximum of electricity that can generate the wind farm of our infrastructure (also see 2.Size calculation).

$\mathbf{Pred\_PPA} \in \mathbb{R}_+^{N_h}$ : represents the predicted quantity of electricity generated by the PPA per hour for the whole fix period ($168 = 7 \times 24$).

$\mathbf{Pred\_Price} \in \mathbb{R}_+^{N_h}$ : represents the predicted prices of electricity on the market per hour for the whole fix period.

**Algorithm 1** First method for prediction

**Require:** $N_w$, $N_h$, FP, DE_solar_profile, DE_wind_profile, DE_Price, DE_Year_index, DE_Week_index, Solar_capacity, Wind_capacity

**Ensure:** Pred_PPA, Pred_Price

1: week_index_list $\leftarrow$ [FP[2] + $i$ for i in 1:Nw-1] ▷ With julia lists begin at 1
2: year_list $\leftarrow$ [0 for i in 1:$N_w$]
3: year_list[1] $\leftarrow$ FP[1]

```
year_list[2 : N_w] ← randint(2015, 2019, N_w − 1)   ▷ pick N_w − 1 integers in ⟦2015, 2019⟧ with replacement
Index_list ← [0 for i in 1:N_h]
j ← 0
for week in week_index_list do
    j ← j + 1
    k ← 0
    for row in 1:length(DE_solar_profile) do   ▷ the last term is included
        if DE_Week_index[row] == week and DE_Year_index[row] == year_list[j] then
            k ← k + 1
            Index_list[168(j − 1) + k] ← row
        end if
    end for
end for
Pred_solar ← Solar_capacity*DE_solar_profile[Index_list]
Pred_wind ← Wind_capacity*DE_wind_profile[Index_list]
Pred_PPA ← Pred_solar + Pred_wind
Pred_Price ← DE_Price[Index_list]
```

### Creation of the optimization problem with constrained levels of stored hydrogen

Here we have a slightly different optimization program that the one presented in the first part :

First, we need to add the constraints for the final level of stored hydrogen of the first $N_w - 1$ weeks. Secondly, we keep only the cost which corresponds to the consumption of electricity from the grid. Finally, we add the notion of efficiency for the storage of electricity in the battery, that means there is still hourly dissipation but there is also a loss of electricity during the transfer of the flow of electricity from the battery.

#### Parameters

$\mathbf{N_w}$ : Number of weeks, it defines the length of the period over which the program will have to optimize.

$\mathbf{N_h} = 168N_w$ : total number of hours of the period

$\mathbf{D} \in \mathbb{N}^{N_h}$ : Customer demand.

$\mathbf{Cost\_PPA} = 0$ : here this electricity is supposed to be free

$\mathbf{Cap\_max\_elec} \in \mathbb{N}$ : maximum quantity of hydrogen that can be produced per hour by the electrolyser (in kg/h)

$\mathbf{Cap\_max\_bat\_flow} \in \mathbb{N}$ : maximum quantity in absolute value of electricity that can go in or out the battery per hour (in MW/h)

$\mathbf{Cap\_max\_bat\_stock} \in \mathbb{N}$ : maximum quantity of electricity that can be stored in the battery (in MW)

$\mathbf{Cap\_max\_tank\_stock} \in \mathbb{N}$ : maximum quantity of hydrogen that can be stored in the tank (in kg)

$\mathbf{efficiency\_elec} \in \mathbb{R}_+$ : Efficiency of the electrolyser (in MWh/kg : it uses electricity to produce hydrogen).

$\mathbf{efficiency\_bat} \in [0, 1]$ : Efficiency of the battery (concerns the losses due to the flows).

$\mathbf{dissipation\_bat} \in [0, 1]$ : coefficient of dissipation of the the electricity contained in the battery per hour.

$\mathbf{Ener\_stored\_bat\_initial} = 0$ : we always start the very first hour of the whole period with empty stock.

$\mathbf{Hydrogen\_stored\_tank\_initial} = 0$ : same consideration.

**Pred_PPA** $\in \mathbb{R}_+^{N_h}$ : predicted quantity of electricity generated by the PPA per hour in MW.

**Pred_Price** $\in \mathbb{R}_+^{N_h}$ : predicted prices on the grid per hour in €/MW.

**hyd_stored_level**(also noted $\mathbf{x_s}$) $\in \mathbb{N}^{N_w-1}$ : corresponds to the value of the minimal stock of hydrogen to have at the end of the first $N_w - 1$ weeks (in kg).

**Variables**

**Ener_PPA** $\in \mathbb{R}_+^{N_h}$ : consumption of electricity generated by the PPA per hour .

**Ener_Market** $\in \mathbb{R}_+^{N_h}$ : consumption of electricity from the grid per hour.

**Ener_bat_flow** $\in \mathbb{R}^{N_h}$ : quantity of electricity going in our out the battery (here we consider that the sign of the flow is positive if we stored electricity in the battery and vice versa).

**Ener_used_by_elec** $\in \mathbb{R}_+^{N_h}$ : total consumption of electricity of the electrolyser per hour (with its efficiency, it gives the amount of hydrogen produced per hour).

**Ener_stored_bat** $\in \mathbb{R}_+^{N_h+1}$ : electricity stored in the battery (we add a component to this vector which represents the final storage at the end of the period, as it was discussed previously).

**Tank_Hyd_flow** $\in \mathbb{R}_+^{N_h}$ : quantity of hydrogen going in our out the tank (same considerations for the sign of this flow).

**Hydrogen_stored_tank** $\in \mathbb{R}_+^{N_h+1}$ : hydrogen stored in the tank.

**Constraints**

1. Initial conditions:
$$\begin{aligned} \text{Ener\_stored\_bat}_1 &= \text{Ener\_stored\_bat\_initial} \\ \text{Hydrogen\_stored\_tank}_1 &= \text{Hydrogen\_stored\_tank\_initial} \end{aligned} \tag{4}$$

2. Demand satisfaction:
$$\forall t \in [\![1; N_h]\!], \quad D_t + \text{Tank\_Hyd\_flow}_t = \frac{\text{Ener\_used\_by\_elec}_t}{\text{efficiency\_elec}} \tag{5}$$

3. Tank flow constraint (we can not use more hydrogen that is currently stored) :
$$\forall t \in [\![1; N_h]\!], \quad -\text{Hydrogen\_stored\_tank}_t \leq \text{Tank\_Hyd\_flow}_t \tag{6}$$

4. Storage in the tank :
$$\forall t \in [\![1; N_h]\!], \quad \text{Hydrogen\_stored\_tank}_{t+1} = \text{Hydrogen\_stored\_tank}_t + \text{Tank\_Hyd\_flow}_t \tag{7}$$

5. Battery flow capacities :
$$\forall t \in [\![1; N_h]\!], \quad |\text{Ener\_bat\_flow}_t| \leq \text{Cap\_max\_bat\_flow} \tag{8}$$

6. Battery flow constraint :
$$\forall t \in [\![1; N_h]\!], \quad -\text{Ener\_stored\_bat}_t \leq \frac{\text{Ener\_bat\_flow}_t}{\text{efficiency\_bat}} \tag{9}$$

7. Storage in the battery:

$$\forall t \in [\![1; N_h]\!], \text{Ener\_stored\_bat}_{t+1} = \text{Ener\_stored\_bat}_t * (1 - \text{dissipation\_bat}) + \text{Ener\_bat\_flow}_t * \text{efficiency\_bat} \tag{10}$$

8. Battery and tank storage capacities:

$$\begin{aligned} &\forall t \in [\![1; N_h + 1]\!], \quad \text{Ener\_stored\_bat}_t \leq \text{Cap\_max\_bat\_stock} \\ &\forall t \in [\![1; N_h + 1]\!], \quad \text{Hydrogen\_stored\_tank}_t \leq \text{Cap\_max\_tank\_stock} \end{aligned} \tag{11}$$

9. Electrolyser capacity :

$$\forall t \in [\![1; N_h]\!], \quad \text{Ener\_used\_by\_elec}_t + \text{Ener\_bat\_flow}_t = \text{Ener\_PPA}_t + \text{Ener\_market}_t \tag{12}$$

10. Energy used by electrolyser:

$$\forall t \in [\![1; N_h]\!], \quad \frac{\text{Ener\_used\_by\_elec}_t}{\text{efficiency\_elec}} \leq \text{Cap\_max\_elec} \tag{13}$$

11. Final levels of stored hydrogen :

$$\forall w \in [\![1; N_w - 1]\!], \quad \text{hyd\_stored\_level}_w \leq \text{Hydrogen\_stored\_tank}_{168w+1} \tag{14}$$

**Objective function**

$$f(\text{Ener\_Market}) = \sum_{t=1}^{N_h} \text{Pred\_Price}_t \text{Ener\_Market}_t \tag{15}$$

With this construction of the optimization problem, we can finally create our subroutine for the heuristic.

**Algorithm 2** Subroutine of the heuristic

**Require:** Parameters of the precedent optimization problem (particularly $x_s$), Pred_PPA, Pred_price, Real_PPA, Real_Price

**Ensure:** Real_Cost

1: Opt_Ener_PPA($x_s$) ← Ener_PPA($x_s$) ▷ solution of the optimization problem using all the inputs
2: Opt_Ener_Market($x_s$) ← Ener_Market($x_s$) ▷ same origin
3: Real_Cost ← $F$(Real_PPA, Real_price,Opt_PPA($x_s$),Opt_Ener_Market($x_s$)) ▷ With F for the transposition of the cost

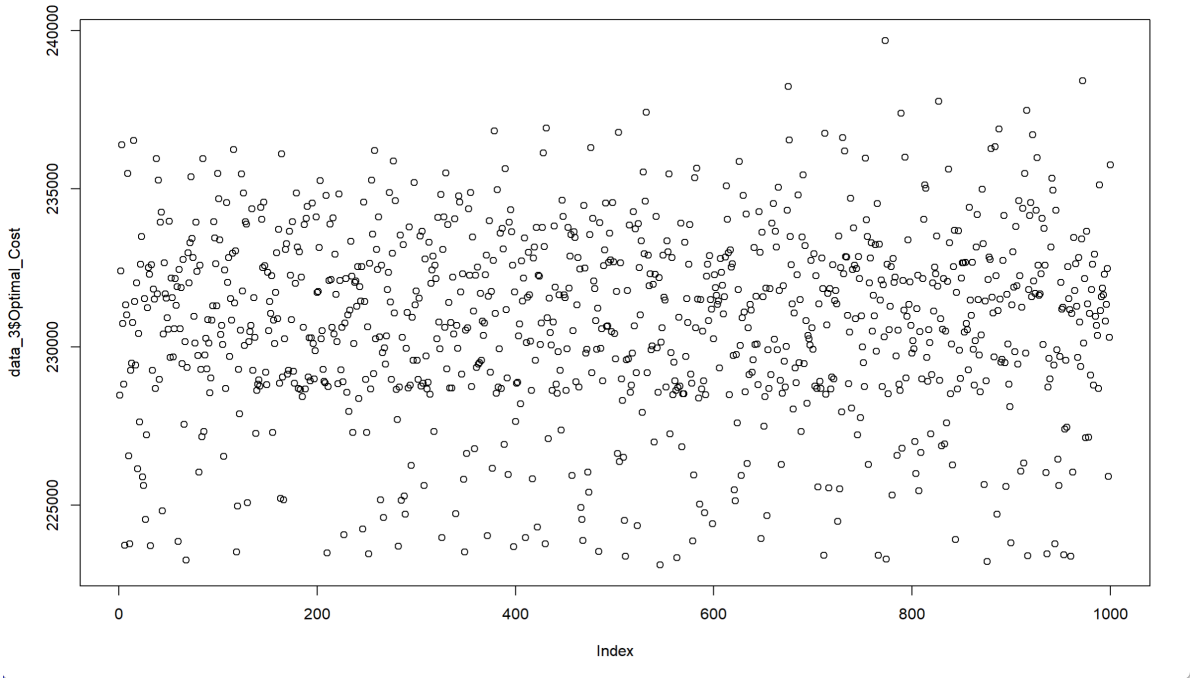
The third part will be described in the next section, but now we test our subroutine with random constrained final levels of stored hydrogen, to have a first look on how the real objective function, that is the function of transposition of the cost, behaves.

## 6.3 First look at the generation of random real costs

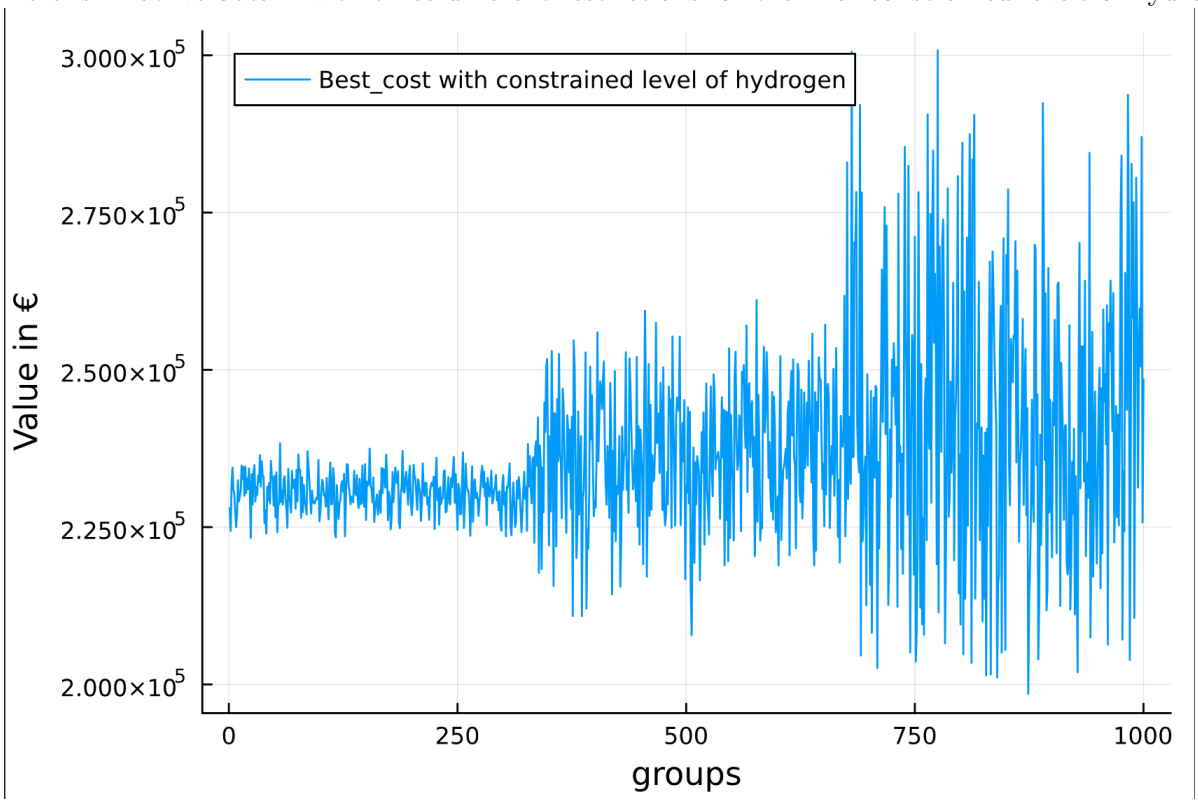
First we look at the optimal cost with random-generated constrained final level of hydrogen, it seems that the different optimal costs do not change a lot, like maybe one percent of difference. This would mean that the different stock of hydrogen do not have a large impact on the cost. This is actually due do the fact that the constrained levels of hydrogen were too low compared to the optimal case where we did not have these constraints, that means these constraints are already verified in the optimal case. Indeed as noticed previously the optimizer tends to saturate the capacities of storage around the middle of the period, and then it lets it decrease until the end. This is a first observation, so not necessarily true for all cases, it actually means that the range of values that the constrained levels of hydrogen can take needs to be large. So instead of restricting

the random levels to be between 0 and 30% of the maximum capacity of storage of the tank, we let it go until 100% while generating more simulations.

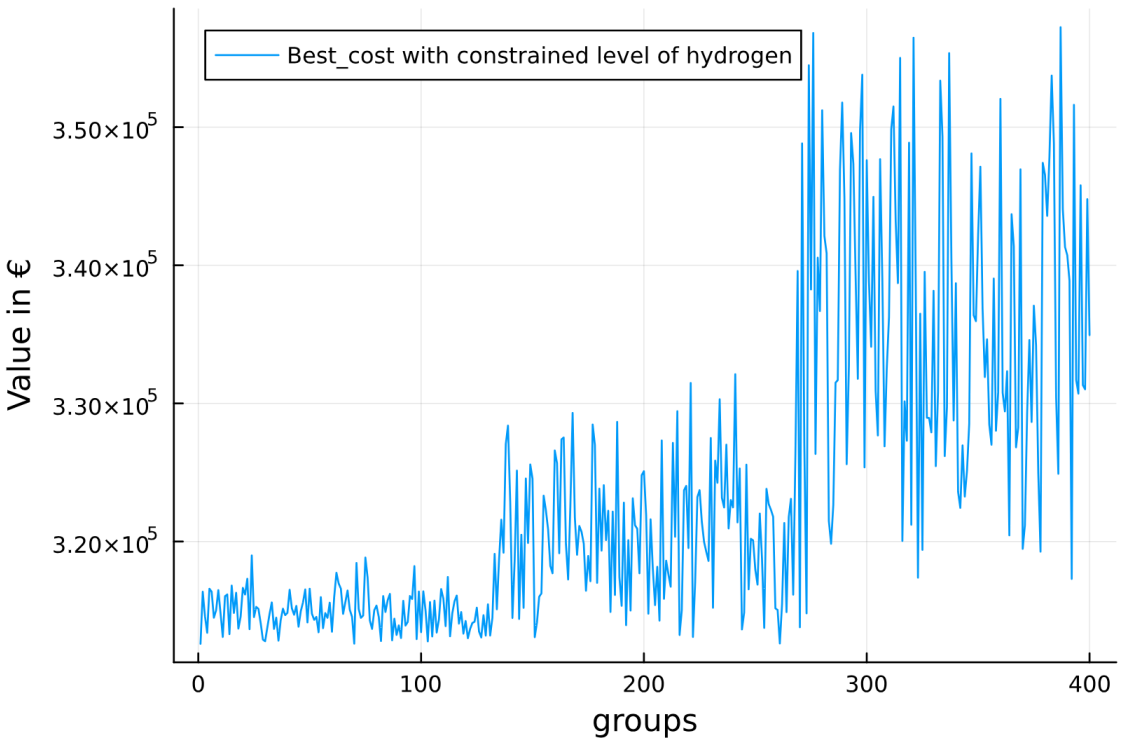
Here is what we obtain with a first set of predicted data with restricting the range of the final constrained levels of hydrogen : it cannot be higher than 80% of the maximum storage capacity of the tank:




Here is what we obtain with three different restrictions for the final constrained levels of hydrogen :



Here is what we obtain with a different set of predicted data:




These graphics are quite interesting, particularly the two last ones. Indeed the first one corresponds to random evaluations of the function of the transposition of costs, that means we used for each evaluation $N_w - 1$ random final constrained levels of hydrogen, and then computed the cost obtained with the real data, namely the transposed cost of the predicted production plan. So here with the first graphic we restricted a bit the range of levels of hydrogen, that means when the final constrained levels of hydrogen were randomly generated, they could take value between 0 and 80% of the maximum capacity of the tank.

What is very interesting with the first graphic is that the first list of constrained levels is initialized to 0, that means we compute the solution of the linear optimization program without adding the level constrained, and therefore it allows to compute the transposed cost that is obtained without a policy for the levels of hydrogen. This will consequently allow to analyse the interest of constraining these levels.

In fact, we observe through the first graphic that some evaluations with non-zero constrained levels have a lower value than the first evaluation that corresponds to the zero case. So with these predicted data constraining the levels has a positive impact on the diminution of the transposed cost, which is a good sign for the interest of the stock constraint method.

The second graphic which presents random evaluations of the transposed function with the same predicted data but with 3 different ranges of stock constraints : 10%, 50%, and 100% of the maximum storage capacity of the tank. It seems with this graphic than greater constraints can allow to reach better transposed cost, that means if we force a high level of hydrogen to reach at the end of a certain week or even at the end of several weeks, then this policy leads to a lower transposed cost.

But the third graphic contradicts the previous observation, the lowest transposed cost is obtained with the null storage constraint policy, and the greater the constraint level can get, the greater the transposed cost is.

The difference here comes from the predicted data, two different predictions were used for the energy and the price. Since for these evaluations of the transposed function we used the method consisting in picking randomly weeks among years as previously described in section 7.2 with the algorithm 2, that means here that for these two graphic two different sets of random weeks were used.

This would mean that the quality of the prediction could have a significant impact on the result of our policy with the final level of hydrogen. To be more precise, there are several ways to observe the quality of the predictions, one consists in using the true data and a certain loss function, this will be develop later. Indeed, we cannot just compute the gap between the predictions an the real data, because although having differences can lead to a non-zero difference between the transposed cost and the real optimal cost : the transposed cost represents the cost of energy needed to apply the predicted optimal production plan, so if the prediction were perfect, then thanks to the determinism of the optimization program we will get the same cost of energy than for the real case and so a difference between these two costs which would be zero), the location in time of these difference could have also an impact. For example having 3 dispersed ill-predicted hours could be actually better than having two consecutive all-predicted error with the same amount of difference with the real data. At this point, this is only an hypothesis, and not an observation.

Another way to study the quality of the prediction, which would be more practical and relieved from the previous hypothesis is to compute the absolute difference between the real optimal cost and the transposed cost obtained with the predicted data and without constraining the levels of hydrogen.

Nevertheless, looking at the gaps between the predicted data and the real data can also be interesting. Indeed, the more we reduce these gaps, the less the difference between the real optimal cost and the transposed cost will be (this evolution definitely may not be only decreasing, in fact reducing this difference means having more accurate predicted data, which are therefore different predicted data, which can therefore lead to a different predicted production plan, and so to a different transposed cost, which is not necessarily closer to the real optimal cost than the previous transposed cost).

## 6.4 Study of the accuracy of the predictions and its impact on the transposed cost

The previous remark can lead to another way of predicting data. Indeed we can introduce a new loss function, which takes other argument than just the real data and the predicted data:

$$\text{loss}(\text{Real\_cost}(\text{Y}), \text{Transposed\_cost}(\hat{Y})) \tag{16}$$

Then the previous remark is not a problem anymore, since more accurate data in the sense of the previous loss function means lower difference between the real cost and the transposed cost.

This method could be a complementary of the policy of constraining the final levels of hydrogen. The issue is that although the idea here is to reduce the difference between the real cost and the transposed cost by constraining the final levels of hydrogen, we saw that having different predictions can lead to completely different impacts of the constraint policy, this is why the way the data are predicted should not be neglected.

Nonetheless the two problems can be studied separately. Indeed, we can use different predicted data and just compare their impact on the impact of constraining the final levels of hydrogen on the transposed cost, meaning that we would search whether having a non-zero policy can lead to a lower transposed cost. These are two separate problem, because although having a lower transposed cost means having a lower difference between the transposed cost and the real cost, since the real cost is a lower bound for the transposed cost, but with some predictions every non-zero policy will only increase the transposed cost, and therefore although these predictions lead to a lower difference between the costs it does not allow to study the impact of the non-zeros policies.

# 7 Heuristic

A heuristic is a practical problem-solving approach or technique that aims to find a good enough solution, even if it is not guaranteed to be optimal. Here are some key points about heuristics:

**Purpose:**
Heuristics are used when finding an exact solution is computationally expensive, time-consuming, or practically infeasible. They provide a way to quickly arrive at a satisfactory result.

**Characteristics:**
- Approximation: Heuristics sacrifice optimality for efficiency.
- Rule of Thumb: They often rely on rules of thumb, intuition, or common sense.
- Domain-Specific: Heuristics are tailored to specific problem domains.

**Examples:**
- Greedy Algorithm: Select the best available option at each step without considering the global impact.
- Hill Climbing: Continuously move toward better solutions by making small adjustments.
- Simulated Annealing: Inspired by annealing in metallurgy, it explores the solution space probabilistically.
- Genetic Algorithms: Mimic natural selection to evolve solutions over generations.

So even if heuristics may not guarantee the best outcome, they often provide practical and efficient solutions in real-world scenarios.

Here we decide to use the simulated annealing, since we have no real idea of how our function behaves.

## 7.1 Simulated annealing

main source : https://enac.hal.science/hal-01887543/document

To quote this paper : Daniel Delahaye, Supatcha Chaimatanan, Marcel Mongeau. Simulated annealing: From basics to applications. Gendreau, Michel; Potvin, Jean-Yves. Handbook of Metaheuristics, 272, Springer, pp.1 35.ISBN 978-3-319-91085-7, 2019, International Series in Operations Research & Management Science (ISOR), 978-3-319-91086-4. 10.1007/978-3-319-91086-4_1. hal-01887543

Simulated annealing is an heuristic which is very helpful in the case where the computation of a value of the objective function is costly in time and in memory, which is our case since since the solver takes 0.30 seconds in average to compute a solution which will be transposed into the reality case. This algorithms relies a lot on probability, it is kind of a mix between evaluating the objective function at random points, and going in the direction which lowers the value of the objective function. The idea is to let randomness decide where the function should go. If the function goes in a better direction, that is the argument does decrease the value of the function, then the function goes that way for sure. But if it goes to a worse direction, then we toss a coin to know if f will actually have to go in that worse direction, hopping it will lead it to a better minimum later. Indeed the idea of the allowance to accept worse direction is made for escaping local minima.

But at some point we could get to the global minimum and then we would like that the possibility to accept a worse direction to be very low. So the idea is to have a criteria of acceptance,i.e., a probability, that decrease over time and with the gap with the previous solution. So even if we can accept a worse direction, the idea is not to accept any worse direction, we do not want our objective function to take tremendous values. These idea are summarized within the Metropolis criterion with f our objective function :

$$\mathcal{P}(\text{"Accept new solution j"}|\text{current solution} = i) = \begin{cases} 1 & \text{if } f(j) < f(i) \\ e^{\left(\frac{f(i)-f(j)}{T_k}\right)} & \text{otherwise} \end{cases} \tag{17}$$

The first concerns are about the choice of the different parameters and the study of them.

Indeed, there are the initial temperature, the coefficient of decreasing of the temperature, the number of iteration per group, the notion of neighbourhood (it could be associated with the variance of a Gaussian centered at the solution point) and finally the initial solution to begin the heuristic.

The policy of acceptance is actually a crucial point of this algorithm, if we reduce the temperature to soon, that means the probability to accept a worse solution is reduced, we could get stuck quickly in a local minimum, but if we take too much time to decrease this temperature we would not converge and could miss the global minimum.

The notion of neighborhood is also quite important since at each iteration we want to test a new solution which is "close" to the most recent accepted one. But we have an argument of dimension 4, so we could update up to four different components. And the size of the space which was visited during this procedure highly depends on the size of a neighborhood. Indeed let us imagine with have a restricted number of iterations as the stopping condition of our heuristic, we begin the algorithm at an initial condition, we compute the value of the objective function at this initial solution, ask the metropolis criterion if we keep it compared to a very high initial score (indeed we do need the first solution to be accepted), and then we compute a new solution in the neighborhood of the first solution. If the neighborhood was too small, we will stay very close to this initial solution and so on. So after the set number of iterations had been reached, we would not have been able to visit the whole space, only a very restricted part around the initial solution, therefore we could have missed the global minimum, or at least a much better final score.

Consequently, the initial solution does also have a major impact on the final solution of the heuristic depending on the stopping condition.

The ideas here will be to test different values for each parameters, different initial solutions and different modeling of the neighborhood, and to compare them in terms of best minimum values and speed of convergence to this value.

## 7.2 Settings for the tests of the heuristic

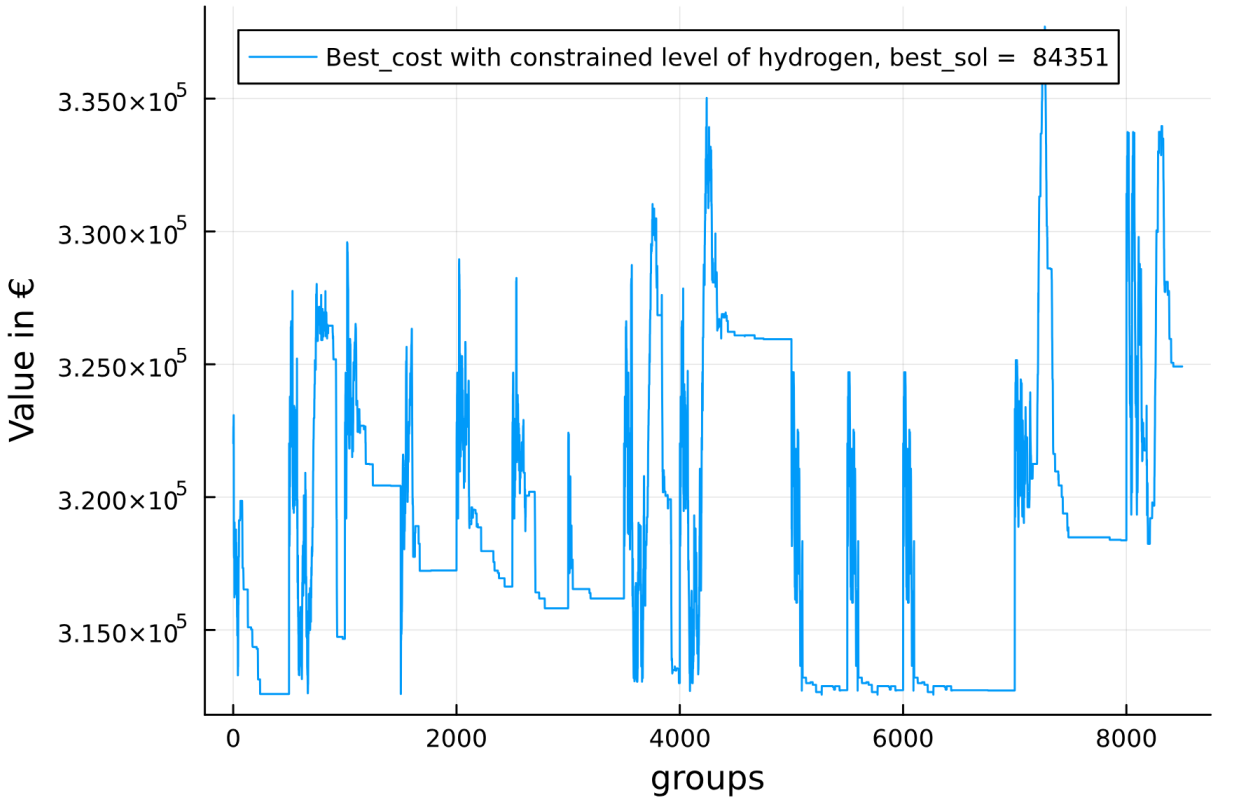
First as explained previously, we need to fix a date, to predict the available energies and prices per hour, then we need to first evaluate the precision of our predictions (since we have the real data), then we need to launch our heuristic with our set of parameters and our predicted data, and see which constrained levels of hydrogen will allow to get to the lowest real cost ,as fast as possible, with our predicted data.

So here the precision of our predictions can have a real impact on the behaviour of our algorithm. Therefor we will have to use better prediction that picking weeks among years.

We also want to ensure reproducibility with our tests, indeed to study how our heuristic works it is better to fix our prediction and change its parameters at first, then obviously we will test our heuristic with other predictions, since in reality we will have different weeks with therefore different prediction, and so the errors will not be at the same hours.

To assure this reproducibility of the tests we will use the seed in Julia, which allows to generate random number while keeping a trace on them.

Here is a graphic showing iterations with the heuristic and different sets of parameters :




**Parameters and variables of the heuristic**

$\boldsymbol{\alpha \in (0,1)}$ : the decreasing coefficient.

$\boldsymbol{S_p} \in \mathbb{N}$ : the size of a patch

$\boldsymbol{\theta \in \{0,1\}}$ : with $\boldsymbol{\theta = 0}$ if the last time split is constrained and $\boldsymbol{\theta = 1}$ otherwise.

$\boldsymbol{H\_ini\_lev} \in \mathbb{N}^{\boldsymbol{N_w}}$ : initial solution

$\boldsymbol{T\_ini} \in \mathbb{R}^+$ : initial temperature

**Var** : variance of the Gaussian law used to model the notion of neighborhood.

Here is the pseudo-code that allows to generate random solution in the Gaussian neighborhood of a previous solution, and the pseudo-code for the Metropois Criterion:

**Algorithm 3** Gaussian neighborhood

**Require:** H_lev, Var
**Ensure:** New_H_lev
  $r \leftarrow$ randint$(1, N_w)$ ▷ a random integer $\in [\![1; N_w]\!]$
  r_list $\leftarrow$ randint$(1, N_w, r)$ ▷ a list of r random integers $\in [\![1; N_w]\!]$, without replacement
  **for** i in r_list **do**
    H_lev[i] = $\lfloor$randnorm$(H_\text{lev}[i], \text{Var})\rfloor$ ▷ Gaussian integer centered at the previous value
  **end for**
  New_H_lev = H_lev

**Algorithm 4** Metropolis Criterion

**Require:** T, $\Delta$
**Ensure:** Decision $\in$ {True, False}
**if** $\Delta < 0$ **then**
    Decision = True
**else**
    $p = \text{rand}(0, 1)$
    **if** $(p < \exp(-\frac{\Delta}{T}))$ **then**
        Decision = True
    **else**
        Decision = False
    **end if**
**end if**

Finally here is the pseudo-code that represents our heuristic "Simulated annealing" :

**Algorithm 5** Simulated annealing

**Require:** T_ini, $S_p$, $\theta$, H_ini_lev, $\alpha$, $N_{\text{iter}}$, Var, $P_{ar}$, Pred_PPA, Pred_Price, Real_PPA, Real_Price
**Ensure:** H_opt $\in \mathbb{N}^{N_w}$
Current_cost $\leftarrow$ Algorithm_2($P_{ar}$, H_ini, Pred_PPA, Pred_Price, Real_PPA, Real_Price) ▷ here H_ini is equivalent to the previous $x_s$
Current_H_lev $\leftarrow$ H_ini
**for** i in 1:$N_{\text{iter}}$ **do**
    **for** k in 1:$S_p$ **do**
        H_lev $\leftarrow$ Algorithm_3(Current_H_lev,Var)
        new_cost $\leftarrow$ Algorithm_2($P_{ar}$, H_lev, Pred_PPA, Pred_Price, Real_PPA, Real_Price)
        $\Delta \leftarrow$ new_cost - Current_cost
        **if** Algorithm_4(T,$\Delta$) **then**
            current_cost $\leftarrow$ new_cost
            Current_H_lev $\leftarrow$ H_lev
        **end if**
    **end for**
    T $\leftarrow (1 - \alpha)T$
**end for**
H_opt $\leftarrow$ Current_H_lev

# 8 Project regression function

## 8.1 Setting of the regression context

This final part allows to finalize the object of the project, namely the impact of constraining the final levels of hydrogen to get the lowest transposed cost.

The context is, we have a fixed period, let us say it is the last day of the year, we want to elaborate an optimal production plan for the next 5 weeks. To do so, we have the parameters of the infrastructure such as the capacities, the efficiencies and the dissipation coefficient of the battery. In this study we consider that these parameters are fixed, so they are not taken into account in the regression problem, contrary to the data used for the prediction of solar and wind profiles and the prices of electricity on the grid. Indeed, since they are times series, each day we can add new data about solar and wind profiles and prices of electricity on the grid.

These parameters and data, associated with the fixed period, namely the date of the beginning of the production plan and the total number of time splits (here we chose weeks) over which we want to construct this production plan, represent the inputs of the regression problem.

The response variable is the vector of the values of the constraints for the final levels of stored hydrogen over the fixed time period (here the last time-split level can be constrained, but in the previous part we did not study that case).

Here are the first considerations. There is possibly different ways to consider this regression problem. We could say that there are only one input, which is the date of the first day. Thus the regression function will depend on the parameters of the infrastructure, on the available data and on the total number of time splits over which the production plan is elaborated, denoted $\boldsymbol{N_T} \in \mathbb{N}$ (previously denoted $\boldsymbol{N_w}$, since the time split was weekly). So with this modelling, let us denote $\boldsymbol{F_P} \in \mathbb{N}^{\mathbf{2}}$, it contains the year and the index of the first time split of this fixed year. We also denote by $\boldsymbol{D_a}$ the data, by $\boldsymbol{P_r}$ the parameters of the infrastructure and by $\boldsymbol{f_{D_a,P_r,N_T,\theta}}$ the regression function. We denote by $\boldsymbol{Y} \in \mathbb{N}^{\boldsymbol{N_T}}$, with $\boldsymbol{\theta = 0}$ if the last time split is constrained ($\boldsymbol{Y_{N_T} \neq 0}$), and $\boldsymbol{\theta = 1}$ otherwise (so previously we had $\boldsymbol{\theta = 1}$). Therefore we have the following regression relation :

$$\boldsymbol{Y = f_{D_a,P_r,N_T,\theta}(F_P)} \tag{18}$$

Now the goal is to estimate $\boldsymbol{f_{D_a,P_r,N_T,\theta}}$, so necessarily with this modelling the data, the parameters and the total number of time splits have to be fixed, otherwise this would lead to another regression function to estimate. Considering the data as fixed is not a problem despite of the fact that each day new data are added, since to make sense the fixed period is always considered in the future. That means if we want to elaborate a production plan for the first five weeks of the year, we fix all the data and parameters the week before, namely the last week of the previous year in this case, and we estimate our regression function. The problem with this modelling is that each time we want to elaborate a new production plan we need to estimate a new regression function since the previous one depended on previous data and now there are new data.

By the previous remark, another idea would be to consider another regression function, which will be refferd in this section as the second regression function. Let us say, we have a large enough dataset for the wind and solar profiles and the prices of electricity on the grid, the regression function will therefore depend only on the data contained in this dataset, on the modelling of the problem (partly seen through $\boldsymbol{\theta}$) the parameters of the infrastructure and on the total number of time splits.

The new data will now be used as variables along the set period for the second regression function. We need to fix the amount of these new data that can be relevant for a fixed period. That means, let us say we estimated the second regression function with data from two years ago. Thus we have two years of new data to elaborate our production plan, but this does not mean that we have to use all these data to find the best constrained final levels of hydrogen for the fixed period, only the the two previous weeks could be useful, since the regression function was estimated with a "large enough" dataset. This consideration is directly linked to the study of times series, especially with ARIMA model, which conveys the exact same idea, namely : how many previous information do we need to have accurate predictions ?

Let us denote by $\boldsymbol{N_{res}}$, the number of previous data we use as input with our second regression function (it could be for example the total number of necessary hours). This number is directly linked to the model of

prediction used for the wind and solar profiles and the prices of electricity on the grid. So it can be estimated outside the estimation of the regression function, but not really outside the regression problem. Indeed, the amount of data the model uses to predict the solar and wind profiles and the prices of electricity on the grid is part of the regression modelling, for example the first regression function depends on the data, but conceptually can deal with different models of prediction (indeed as presented before we could have different loss function for the accuracy of the prediction, thus leading to different models of prediction).

Let us now denote by $\boldsymbol{D_{a\_\text{new}}(F_P)}$ the new data, and by $\boldsymbol{D_{a\_\text{new\_res}}(F_P)}$ their restriction used as input for the second regression function. The second regression function depends on $\boldsymbol{N_{res}}$, so a different number would lead to a different regression function, therefore this quantity has to be constant and not to depend on the fixed period, contrary to the new data which depend by definition on the fixed period (if we consider that the fixed period corresponds to the current period).

So here we give the regression relation verified by the second regression function denoted by $\boldsymbol{f_{D_a,P_r,N_T,\theta,N_{\text{res}}}}$ :

$$\boldsymbol{Y = f_{D_a,P_r,N_T,\theta,N_{\text{res}}}(D_{a\_\text{new\_res}}(F_P))} \tag{19}$$

The efficiency of the second regression function, compared to the first one, relies more on the accuracy of the predictions. It could be seen as working with an old dataset to elaborate new production plan, hence the interest of having a old "big enough dataset". Consequently, in the following we will work with the second regression function considering that the dataset is "big enough".

## 8.2 The regression function

As explained above we now work with the second regression function. We will here develop its construction and its links to its dependencies.

First to estimate this regression function we need to compute manually couples $\boldsymbol{(Y, D_{a\_\text{new\_res}}(F_P))}$ that will be used for the estimation as observations. To do so, we first need to precise that here the term new data does not make sense, since we use a fixed dataset for the computation of the observations. Therefore here, only for the computation of the observation we should replace $\boldsymbol{(Y, D_{a\_\text{new\_res}}(F_P))}$ by $\boldsymbol{(Y, D_{a\_\text{res}}(F_P))}$.

Then, to compute these couples, we need to have a model of prediction, which will be trained on the initial dataset (solar and wind profiles and prices of electricity on the grid) hence the first reason for the dependency of the regression function on the dataset $\boldsymbol{D_a}$. The second reason comes from the fact that for the computation of the transposed cost we need the real data and the predicted data of the fixed period. But here, there is actually a major consideration concerning the training of the prediction model, since to compute the observations we need to evaluate this prediction model at "new" data, but as explained above here there is no "new" data. That means if here we train and validate the prediction model on the whole dataset, and then we evaluate it to generate predicted data which actually correspond to data already present in the dataset, then we will get incredibly accurate prediction, since the model was trained and validated on them. It is the exact same consideration that leads to have a training set and a validation set to elaborate prediction model that can generalize well. But here it is bit more subtle, since here the most important is the accuracy of the regression function, which is partly linked to the accuracy of the prediction model (it is also linked to the efficiency of the heuristic). To be clear, if the prediction model is really bad at generalizing, then the regression function will also be at it (assuming that the efficiency of the heuristic does not lead to an opposite result, which would be unlucky...), since the regression function takes as input new data that should be passed to the prediction model to get predicted data for the fixed period.

So to compute these observations, we need to split the dataset into two sets, one which will be used to elaborate the prediction model (so this set will probably be cut into three smaller sets, one for the training, one

for the validation, and the last one the final test phase), and the other which will be used as input data for the prediction model once trained, validated and tested, to compute the observed couples $(\boldsymbol{Y}, \boldsymbol{D_{a\_res}(F_P)})$. Before in the subsection 7.2, the algorithm use to predict data did not have these issues, since the modelling was not the same. Indeed it was not a trainable model, since it randomly picked weeks among years from the dataset. Actually having an untrainable prediction model can be useful if the dataset is too small, and therefore cannot be split into two "large enough" sets. Nethertheless, the modelling of the randomly-picked-weeks cannot really be used with the modelling of this regression function, since it only needs the fixed period among the two inputs of the regression function to output predicted data. So that would mean that only the time slot index (here it is weekly) can have an impact on the response variable, in other words the production plan should have final constrained levels of hydrogen whose values only depend on the time slot index of the considered fixed period.

Having now the predicted data for a chosen fixed period $\boldsymbol{F_P}$ that were output by a specified prediction model, trained with the dedicated part of the initial dataset, and taking $\boldsymbol{D_{a\_res}(F_P)}$ (so here necessarily $\boldsymbol{D_{a\_res}(F_P)}$ belongs to the second part of the initial dataset) as input, we can run our heuristic, which will iterate on different final constrained levels of hydrogen to get the lowest transposed cost. Since this is an heuristic, we need to define a stopping condition, here it would be the number of iterations (it could be something different, such as if the heuristic gets stuck too long). Once the heuristic is done, we select the best group of constraints which will represent the Y for this observation.

We repeat the previous part, by choosing new time period $\boldsymbol{F_P}$ and their associated "new data" $\boldsymbol{D_{a\_res}(F_P)}$, until we have enough observations to estimate the regression function $\boldsymbol{f_{D_a,P_r,N_T,\theta,N_{res}}}$.

Here is the pseudo code for the algorithm that summarize the creation of the observations that will be used for the estimation , we will denote by $\boldsymbol{N_o}$, the numbers of required output observations, and by $\mathbf{Pred\_Mod}$ our elaborated prediction model which was trained and validated on $\boldsymbol{D_{a\_pred}}$, such that $\boldsymbol{D_a = D_{a\_pred} \cup D_{a\_heuris}}$ and $\boldsymbol{D_{a\_pred} \cap D_{a\_heuris} = \emptyset}$. A simple way to deal with the previous partition of the data, and their selection and restriction, we can vertically concatenate our data. That means our data are ranged with chronological order, the first line correspond to the first hour of a year and the last one correspond to the last hour of another year, which happened after the first one. Let us denote by $|\boldsymbol{D_a}|$, the total number of lines(hours) contained in our dataset, by $\mathbf{prop\_pred}$ the proportion of the dataset used for the training and validation of the prediction model, thus we have : $\boldsymbol{|D_{a\_pred}| = prop\_pred|D_a| \in \mathbb{N}}$ and $\boldsymbol{|D_{a\_heuris}| = (1 - prop\_pred)|D_a| \in \mathbb{N}}$. Now we have by choosing to split our dataset such that the first part is dedicated to the prediction model and the second part is dedicated to the heuristic : $\boldsymbol{D_{a\_pred} = D_a[1 : |D_{a\_pred}|]}$ and $\boldsymbol{D_{a\_heuris} = D_a[|D_{a\_pred}|+1 : |D_a|]}$. This is just an easy way split the data, and therefore not necessarily the best one, it will just make the algorithm clearer and more readable. Finally, we denote by **Heuris**, the heuristic that will use the subroutine presented in the subsection 7.2 through the algorithm 2.

**Algorithm 6** Generation of observations

**Require:** $D_{\text{a\_heuris}}$, $P_r$, $N_T$, $\theta$, $N_{\text{res}}$, $N_o$, stop_event
**Ensure:** $Z = (Y, X) \in (\mathbb{N}^{N_T} \times \mathbb{R}^{3\times N_{\text{res}}})^{N_o}$

r_list $\leftarrow$ randint($N_{\text{res}}, |D_{\text{a\_heuris}}| - N_T, N_o$) ▷ a list of $N_o$ random integers $\in [\![$N_res; $|D_{\text{a\_heuris}}| - N_T]\!]$ without replacement
$Z \leftarrow$ []
**for** i in 1:$N_o$ **do**
    Pred_solar_ $\leftarrow$ Pred_Mod($D_{\text{a\_heuris}}$[r_list[$i$] $- N\_res$, r_list[$i$]]$_{|\text{solar}}$)
    Pred_wind $\leftarrow$ Pred_Mod($D_{\text{a\_heuris}}$[r_list[$i$] $- N\_res$, r_list[$i$]]$_{|\text{wind}}$)
    Pred_PPA = solar_capacity $\times$ Pred_solar + wind_capacity $\times$ Pred_wind
    Pred_prices $\leftarrow$ Pred_Mod($D_{\text{a\_heuris}}$[r_list[$i$] $- N\_res$, r_list[$i$]]$_{|\text{prices}}$)
    Real_solar $\leftarrow$ $D_{\text{a\_heuris}}$[r_list[$i$] $+ 1$, r_list[$i$] $+ N_T$]$_{|\text{solar}}$
    Real_wind $\leftarrow$ $D_{\text{a\_heuris}}$[r_list[$i$] $+ 1$, r_list[$i$] $+ N_T$]$_{|\text{wind}}$
    Real_PPA = solar_capacity $\times$ Real_solar + wind_capacity $\times$ Real_wind
    Real_prices $\leftarrow$ $D_{\text{a\_heuris}}$[r_list[$i$] $+ 1$, r_list[$i$] $+ N_T$]$_{|\text{prices}}$
    $Y \leftarrow$ Heuris(Pr, $\theta$, Pred_PPA, Pred_prices, Real_PPA, Real_prices, stop_event)
    $Z[i] \leftarrow (Y,$ Pred_solar, Pred_wind, Pred_prices)
**end for**

Now that we have generated our $N_o$ observations, we can get into the regression area, that means using statistical inferences to get the best regression function in terms of transposed costs.

## 8.3 Estimation of the regression function

Depending on the size of the dataset and on the computation power, we can assume that we have $N_o < 3N_{\text{res}}$, this means that we are in a case of statistics in high dimensions, in other words we have more predictors than observations. So different methods of regression could be testes, for example a decision tree, which allows to select the most significant dimension, or a ridge or lasso regressions, than can also deal with high dimension.