NONPARAMETRIC STATISTICS PROJECT

Academic year 2021-2022

cured by ***Matteo Bollettino, Jaime Enriquez, Madhurii Gatto***

This is the report of the project lead for the Nonparametric Statistics course at Politecnico di Milano.
GitHub repository link: https://github.com/ebjaime/NPS-project

ABSTRACT

During the last few decades, the European Union has made several proposals to stimulate the creation of clean energy inside its country members (*Kyoto protocol, Paris Agreement, European Green Deal, Clean Energy for all Europeans, 2030 Climate and Energy EU objectives*…) with the common objective of having a total renewable energy future. Several projects have been carried out trying to predict whether these objectives are plausible in the near future. A nonparametric approach seemed us like an interesting path which could help analyze the effort made regarding this cause since it permits to obtain consistent results under few conditions. We focused our attention on the *European Green Deal* (2020) that European Commission President Ursula von der Leyen considered *"like man landing on the moon"* for Europe. In fact this pact would make Europe the first continent to achieve Climate Neutrality by 2050 and in particular it enhances Paris objective with the new target of 55% cut of greenhouse gas emissions by 2030 respect to 1990. Given the great importance for Europe that this pact would have if it were respected we have set ourselves the goal of predicting whether this Green Deal target of a 55% cut in greenhouse gas emissions (Ghg) by 2030 respect to 1990 levels is feasible. To do so we analyzed the European renewable transition and forecasted future renewable energies and Ghg values, taking advantage of current and past knowledge of data about production, consumptions and greenhouse gases. Since all energy related information for every EU member is open to the public in the EU's Eurostat portal, and all of its datasets contain clean and reliable information, we deemed this project very interesting and useful to draw further attention to these issues while using the nonparametric techniques learnt. We worked on a variety of time series datasets about European energy consumption and of Ghg emissions in the time window from 1990 to the present (*see References*). To properly predict future renewable energy and Ghg values and to make our models more robust we also used population and Gross Domestic Product (GDP) reliable predictions (*see References*).

DATA DESCRIPTION

To reach our ambitious goal, we gathered a variety of time series datasets about different European energy indicators in the time window from 1990 to the present, available on the EU's Eurostat portal (*see References*). In particular the energy quantities analyzed in our study are:

- Non-renewable Energy Inland Consumption
- Renewable Energy Inland Consumption
- Use of Renewable Energy for Electricity Production
- Total Energy Production Capacity
- Combustible Fuels Energy Production capacity
- Renewable Sources Energy Production capacity
- Greenhouse gas emissions (percentage wrt 1990 value)

We chose not to include nuclear energy neither in renewable nor in non-renewable sources, since this is an argument of debate and furthermore this type of energy was recently (February 2022) labeled as a "transitional energy source". Together with nuclear energy, also natural gas was recently labeled as "transitional", but since the project was started in late 2021, we clearly included it in the non-renewable sources.

In each dataset we have European Countries as sample units (as well as the EU totals), with their relative time series unrolled on the columns.

These datasets required a bit of pre-processing to make them ready for the planned analysis.

For this reason we implemented a short function which cleans the rows and columns which we do not need, removes the observations with some missing data, and converts the measurements to Terajoules, where needed. In fact non-renewable sources consumptions like Fossil Fuels and Oil and Petroleum were initially expressed in thousand tonnes (of oil equivalent), while renewable sources consumptions were already described in Terajoules. Using the respective calorific powers, we unified the units of measurement.

Data regarding electricity production were expressed in Gigawatt-hour, while the ones regarding the electricity capacity were expressed in Megawatt, so no conversion was necessary here.

For our core section of the project, extra datasets were incorporated to the project. In particular, we decided to include data on population and on Gross Domestic Product (GDP) as explanatory variables for our regression models. Past values were obtained from the World Bank portal (*see References*). These were combined with predictions for future values, obtained from the Socio Economic Pathways project. As a brief summary, from their official paper (*see References*): "(SSPs) describe plausible major global developments that together would lead in the future to different challenges for mitigation and adaptation to climate change". 5 possible scenarios (SSP1, SSP2, SSP3, SSP4, SSP5) are presented as possible futures, ranging respectively from a best possible case (clean energy growth, low population growth...) to a catastrophic future (low regard for global environment, high population growth…). From these scenarios, we try to stick as close as possible to reality and choose SSP2, described as the most possible scenario ("[...] moderate challenges of both kinds and is intended to represent a future in which development trends are not extreme in either of the dimensions, […]").

The particular joint dataset combines values of GDP (in billions of current dollars) and population from 1990 to 2009 from the World bank dataset and values from 2010 to 2030 from the SSP2 scenario.

Data from the World bank dataset was not used from 2009 to 2021 because of the possible jump for some countries when joined with the predictions of SSP2 on 2020. Since predictions for the SSPs start from 2010, and no big jump is found with the real data of 2009, it seemed us the most reasonable choice.

A smoothing procedure was done to obtain year-to-year values for the SSP indices, since Population and GDP predictions were made with a periodicity of 5 years (i.e. for 2010, 2015, 2020, 2025...). This transformation can be found in the *preprocessing_SSP.Rmd* script, inside the *src/* folder. To be brief, the smoothing was done with each 5 year prediction as a knot of the model and taking 2nd order basis functions.

The problem of comparing past GDP values with current ones is that inflation is not taken into account (i.e. a conversion should be made for all values to be explained in current US dollars). The SSP GDP predictions are made with the 2005 US dollar value while the World Bank dataset are expressed in current US dollars. The US bureau of labour statistics includes a simple transformation (*see References*) to express 2005 US dollars in current US dollars, which we translate into a function.

ANALYSIS OF CURRENT EUROPEAN RENEWABLE TRANSITION

To properly deal with our objective we thought it was important to analyze, as a starting point, the current European situation and in particular its "Green Transition".
We structured this analysis in 3 important consequent steps to study in detail the European situation.

1. The first target was to **investigate how renewable and non-renewable consumptions are evolving.**

We started by statistically testing with a permutational Manova (using Wilks as permutational test statistic) if the consumptions coming from the three main non-renewable sources (solid fossil fuels, natural gas, oil and petroleum products) follow the same distribution. We can visualise the three consumption means in figure 1.
As we can see from figures 2-3, we reject the null hypothesis (p-value=0) so we can state that at least one of them has a different distribution from the others.

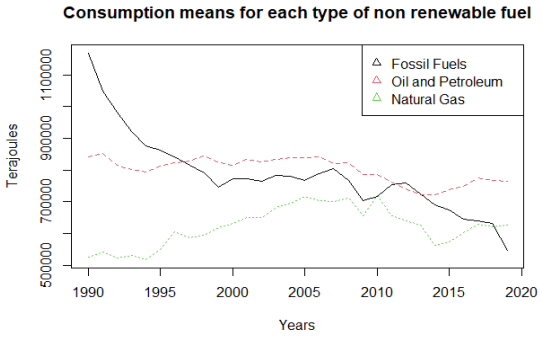

Figure 1

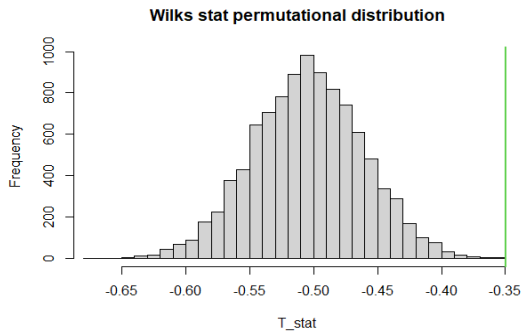

Figure 2

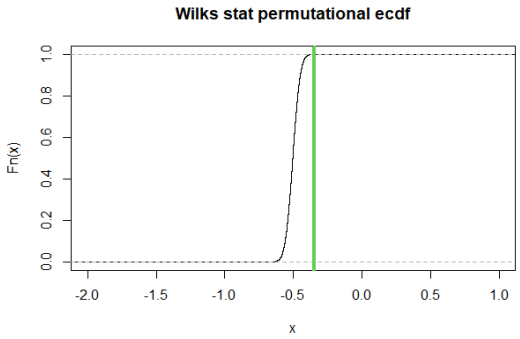

Figure 3

To investigate whether the difference was in their means we proceeded with 3 coupled permutational tests using the squared Euclidean distance (i.e. sum of squared differences) between the two sample mean vectors obtaining: Natural gas vs oil & petroleum p-value = 0.4901, Fossil fuels vs oil & petroleum p-value = 0.7836, Natural gas vs fossil fuels p-value = 0.4618. We never rejected the null hypothesis so we could assume that in average the 3 non-renewable sources are decreasing at the same rate. This result gave us the possibility to compare renewable energies with the average of non-renewable ones. In particular we wanted to know if the renewable consumptions are increasing as fast as the non-renewable consumptions are decreasing.
To do so we performed a permutational test using the same statistics as before so considering the distance between the means of year-to-year variations of consumptions, so we made a test over the first derivatives (Figures 4-5). In particular we changed sign of non-renewable year to year differences to proceed comparing if the two curves are moving at the same rate. Since we do not reject the null hypothesis of same growth rate (p-value = 0.3865) we can effectively conclude that, in average, **the renewable consumptions are increasing as fast as the non-renewable consumptions are decreasing**.

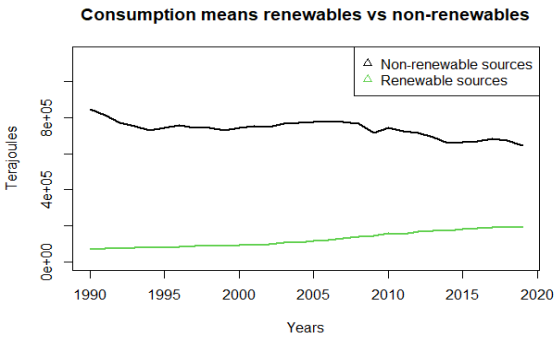

Figure 4

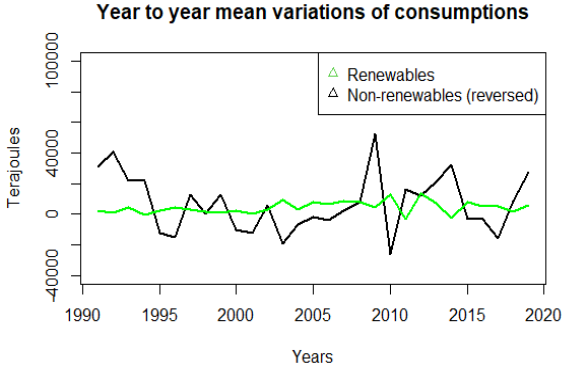

Figure 5

2. The second target was to **investigate how electricity production coming from renewable sources is growing**.

To get a first insight of the evolution of renewable electricity production we plotted the percentage of European electricity that derives from renewable sources and as we can see (Figure 6), it is clearly growing with time. Furthermore we visualized the trends of all EU countries electricity production from the three main renewable sources (Hydro, Solar, Wind) plotting (Figures 7-8) both their values and their yearly variations.

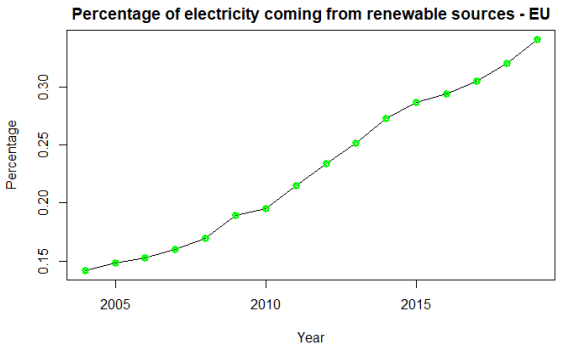

Figure 6

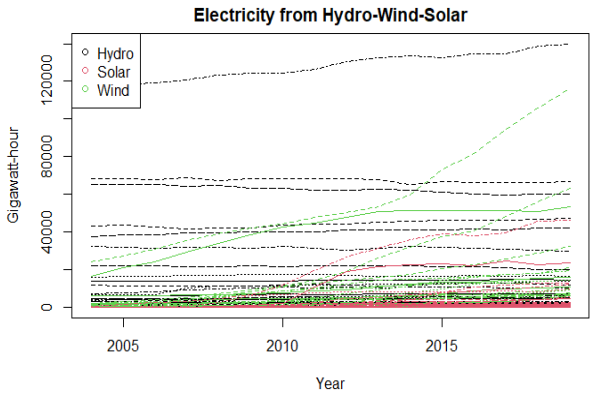

Figure 7

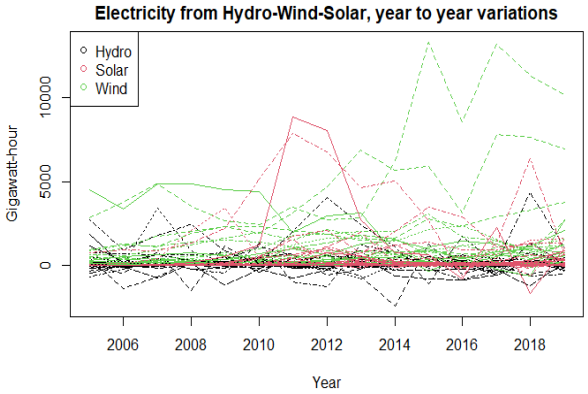

Figure 8

It seemed like Hydro is uniformly more used than the other two and has less variations than the other two, so we proceeded to statistically test this last fact.
To avoid multicollinearity we proceeded with a manova test, using Wilks as permutational test statistic, only on year-to-year variations to investigate whether the increase of electricity production coming from the three main renewable sources (Hydro, Wind, Solar) has the same distribution. From the results (p-value = 0.0139) of the "permutational" manova we can state that at least one of them has different growth distribution from the others.

As observed before, Hydro seems to be the principal source of renewable electricity and the one with less variations, so we proceeded investigating if the two other sources (Wind and Solar) are in average developing at the same rate testing the hypothesis first globally with a permutational test and then, in case of rejection, locally exploiting the Interval-wise Testing procedure, still using their year-to-year variations (Figure 9). In the end we have no statistical evidence (p-value = 0.0615) to state that the two growth distribution are different so local testing was not necessary.

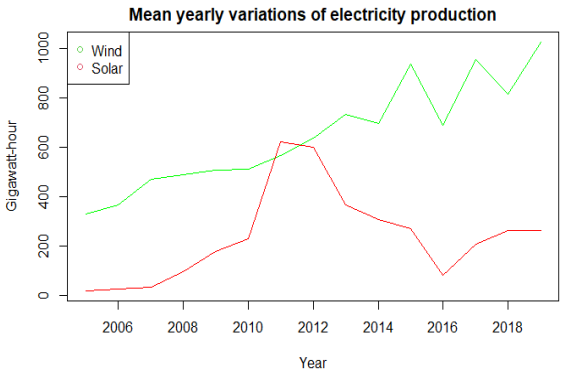

Figure 9

To better understand how Hydro, Wind and Solar usage for electricity production is growing we proceeded with some regression on the amount of electricity coming from each of them. In particular we decided to work with total EU values. We first tried to construct Generalized Additive Models (GAM) using original data and checked the correlation of the residuals using the (partial) autocorrelation function, obtaining high correlation values. This lead us to change strategy in order to tackle this problem in most of the cases: we looked for GAM models to derive the year-to-year variations of each electricity source in function of GDP and population year-to-year variations and their interactions. Each time we searched for the best model by looking at the significance of the covariates and the adjusted R

squared. Each time, to test whether a term was significant or not, we exploited the anova test for model comparison if the residuals were gaussian, while we exploited a permutational anova for model comparison if they were not. The three obtained models are the following:

1. DIFF_EU_WIND ~ β_0 + f(DIFF_EU _GDP) + DIFF_EU _POPULATION + f(DIFF_EU _GDP* DIFF_EU _POPULATION) + ϵ

2. DIFF_EU _HYDRO ~ β_0 + f(EU _GDP) + f(EU _POPULATION) + f(EU _GDP* EU _POPULATION)) + ϵ

3. DIFF_EU _SOLAR ~ β_0 + f(DIFF_EU _GDP) + DIFF_EU _POPULATION + f(DIFF_EU _GDP* DIFF_EU _POPULATION) + ϵ

Where:
EU_variable means the original values of that variable,
DIFF_EU_variable means the year-to-year variations of that variable,
f(variable) means a cubic spline base for that variable.

We then proceeded to plot the obtained regression planes (Figure 10 – Wind, Figure 11 – Hydro, Figure 12 -Solar) along with the three temporal evolutions (Figure 13) obtained with these models, as you can see below.

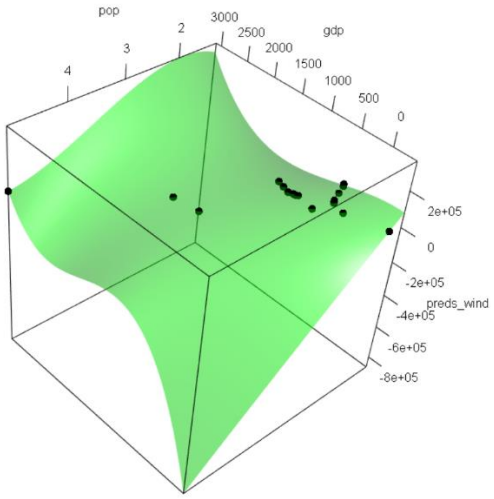

Figure 10

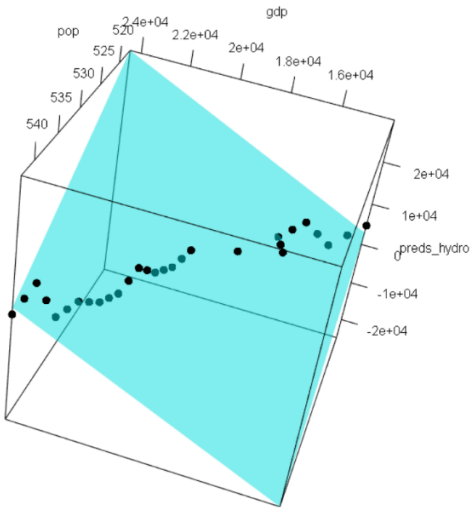

Figure 11

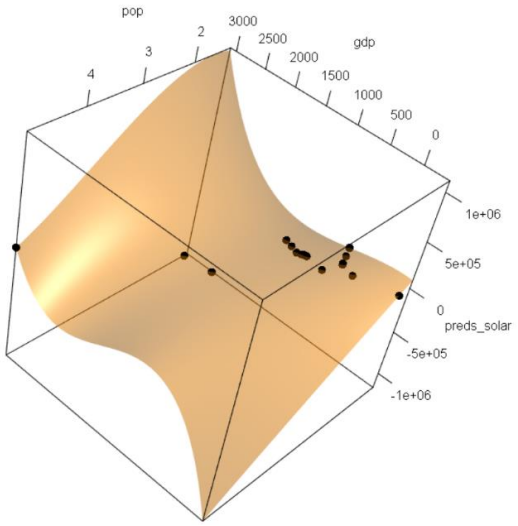

Figure 12

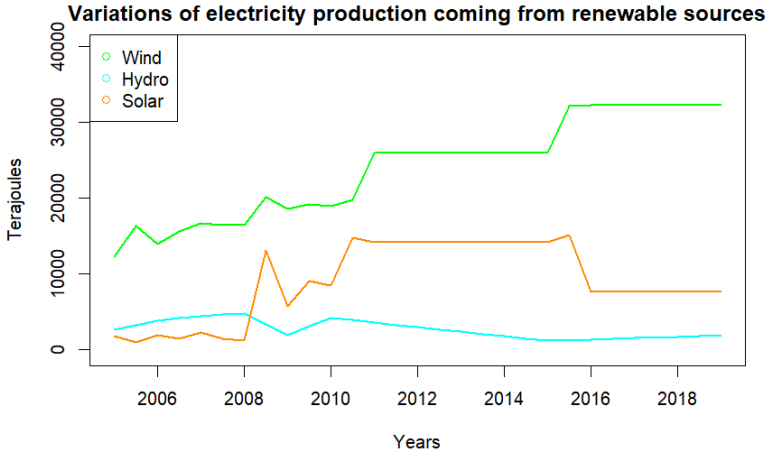

Figure 13

The conclusion of this second step is that **electricity production coming from renewable sources is effectively growing**; in particular wind and solar are developing at the same rate, while Hydro is the most stable (and already developed) source of electricity.

3. The third target was to **analyze the current European renewable energy network**.

One of the most important indicator to analyze whether we are moving towards a "green future" is to investigate whether electricity power coming from renewable sources would be enough to satisfy picks of electricity demand, i.e. understand if the renewable energy network has sufficient production capacity, in terms of daily generated power (Megawatt).

To do so we start investigating what the actual situation is with some graphs. Figure 14 regards the total electricity capacity of all European countries. Germany (light blue line) is the country that seems to have the largest amount of capacity from both renewable and non-renewable sources, as we

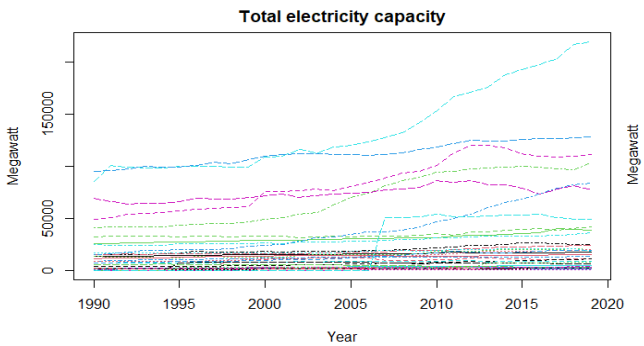

Figure 14

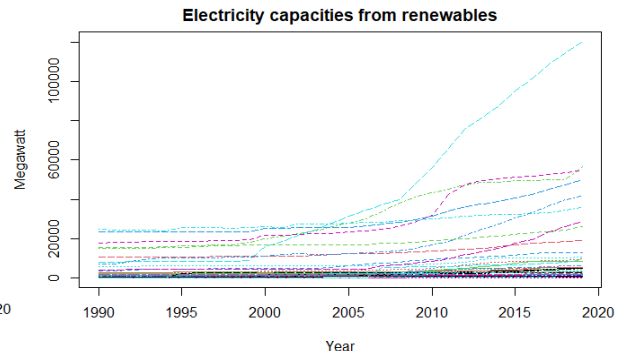

Figure 15

can see also in figures 15 about electricity capacities from renewable sources. This is reasonable since it is one of the biggest european country. To properly view the trend of capacity growth we then decided to use Natural Cubic Splines on total European capacities (Figure 16) because it turned to be better than using a global approach, since it better captures temporal local variations.

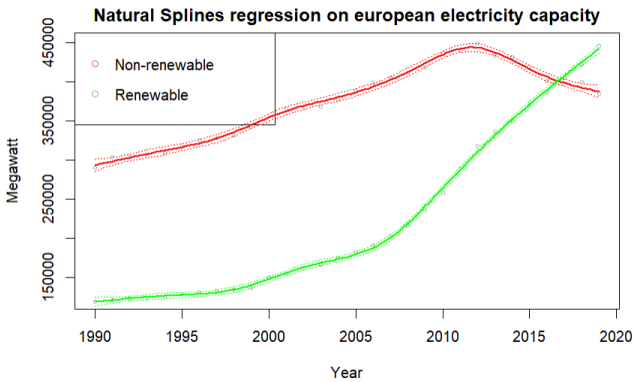

Figure 16

We continued by looking for different development of the renewable network in the four regions of Europe (North, East, South and Centre) (figure 18) with an anova test on 2019 relative capacities. These quantities were computed as the ratio between renewable and total capacities in order to take into account the country dimension. As we can see from both the boxplot (figure 17) and the result of the permutational test (p-value = 0.3230) we have no evidence to reject the hypothesis that the capacity distribution is different in the four european regions.

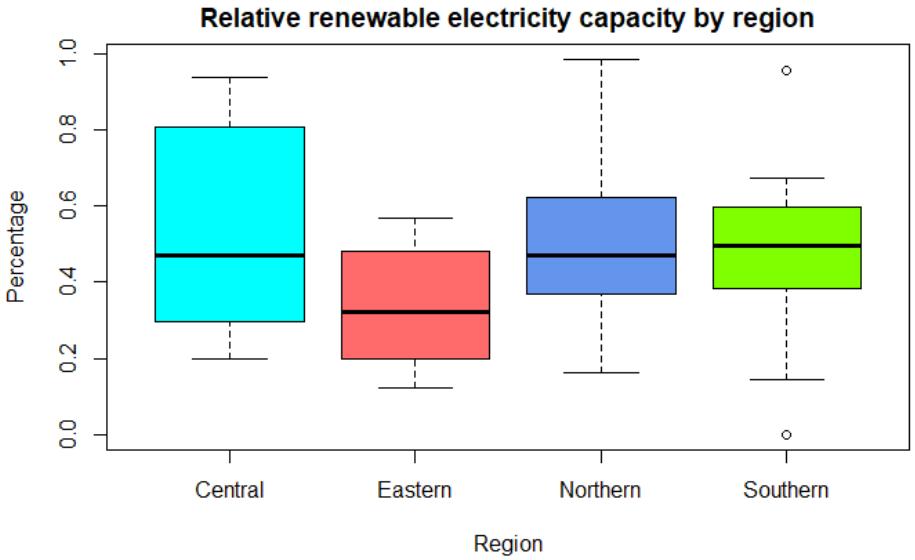

Figure 17

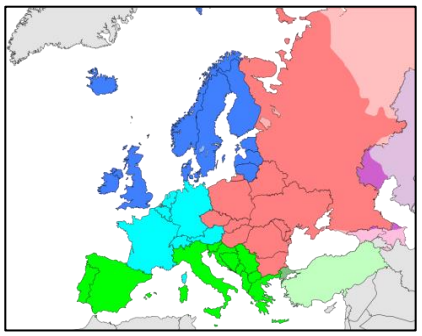

Figure 18

So to conclude this third analysis step, we can state, from the above results, that **electricity capacity coming from renewable sources has surpassed non-renewable one** around 2016 and that this **improvement is uniformly distributed across the four regions of Europe** since there is no evidence to say that a certain region is doing better than the other ones.

So the conclusion of the start point analysis is that:
Europe is effectively undertaking a renewable transition... but is it fast enough?

To answer to this question we are ready to enter in the core of our project that is to construct a robust model to predict future european Greenhouse gas emissions, to properly investigate whether the Green Deal target of a 55% cut in Greenhouse gas emissions by 2030 respect to 1990 levels is feasible.

PREDICTING FUTURE EUROPEAN GREENHOUSE GAS EMISSIONS

With the target to develop a consistent model to predict future Ghg emissions as a function of the other energy indicators we had to previously construct a variety of auxiliary models to predict covariates values to be used in the final model.

Since we are working with time series, where there is a strong temporal dependence, from now on each time we constructed a model we checked the correlation of the residuals using the (partial) autocorrelation function. This is important to understand whether the model is able to capture the relations in original data or whether it is better to perform a first order differentiation to mitigate the natural collinearity problem of these datasets. With this idea in mind, we initially fit the models with original data and then saw that it was necessary to switch to yearly variations. Here we only present the results using year-to-year differences of datasets since they are the more reliable ones.

The first auxiliary models we fit aim to point predict year-to-year differences of the total European consumption from the three non-renewable sources (oil and petroleum products, solid fossil fuels and natural gas) in function of gdp, population and their interaction. In particular we constructed the following 3 models each time searching for the best model by looking at the significance of the covariates and the adjusted R squared. Again, to test whether a term was significant or not, we exploited the anova test for model comparison if the residuals were gaussian, while we exploited a permutational anova for model comparison if they were not. We always started from a GAM model and iteratively quit smooth terms if not even one of the terms in the model was significant.

Models:

1. DIFF_EU_OIL&PETROLEUM_PRODUCTS ~ β_0 + DIFF_EU_GDP + DIFF_EU_POP + ϵ

   The obtained model considers linear contributions both from gdp and population. No interaction term in this case. We present the summary of the model (Summary 1), along with the regression plane (Figure 19) and the temporal evolution plot (Figure 20).

```
Call:
lm(formula = dif_oil_eu ~ dif_gdp_eu + dif_pop_eu, data = vec_oil.diff)

Residuals:
    Min      1Q  Median      3Q     Max
-811448 -420349  110512  351005  652233

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  345845.6   223232.8   1.549   0.1334
dif_gdp_eu      533.4      229.3   2.326   0.0281 *
dif_pop_eu  -472120.3   190346.4  -2.480   0.0199 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 444500 on 26 degrees of freedom
Multiple R-squared:  0.2038,    Adjusted R-squared:  0.1426
F-statistic: 3.328 on 2 and 26 DF,  p-value: 0.05167
```

Summary 1

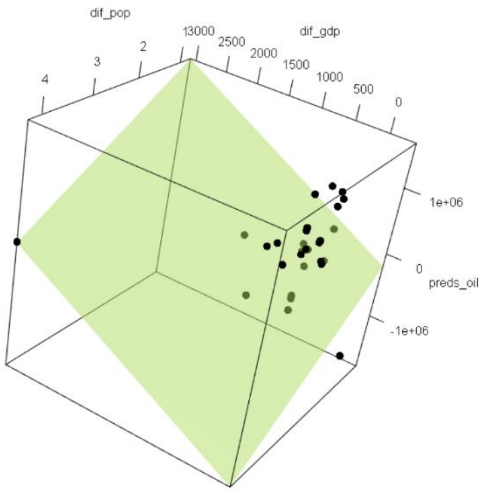

Figure 19

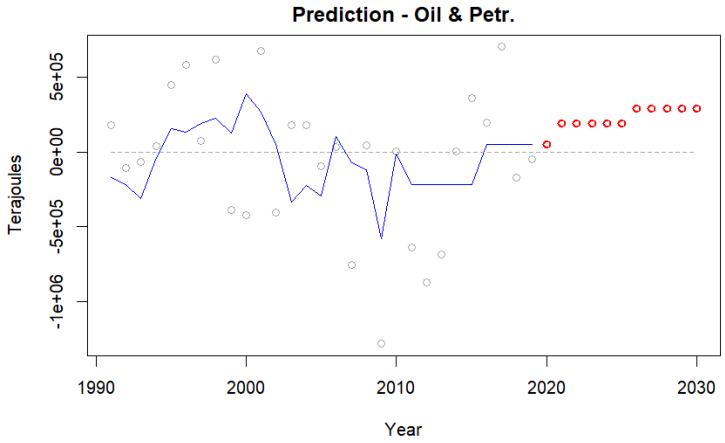

Figure 20

Note that (Summary 1) it is quite strange that we should not reject the null hypothesis of the fundamental test of significance, while we reject the null hypothesis of the singular tests. Therefore we checked the significance of the whole model, exploiting a permutational test. We obtained a p-value equal to 0, so we could reject the null hypothesis of non-significance of the model and confidently proceed with this model. Note also that (Figure 20) the predictions (and also the ones following models) are quite “at steps”. This should be due to the smoothing we had to apply to gdp and population datasets, whose future predictions were available each five years.

2. DIFF_EU_SOLID_FOSSIL_FUELS ~ β_0 + f(DIFF_EU_POP) + ϵ

The obtained model only considers a cubic smooth term on the population. No gdp term and no interaction term. We present the summary (Summary 2) of the model, along with the regression plane (Figure 21) and the temporal evolution plot (Figure 22).

```
Family: gaussian
Link function: identity

Formula:
dif_fossil_eu ~ s(dif_pop_eu, bs = "cr")

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -686495     144163  -4.762 9.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                edf Ref.df     F p-value
s(dif_pop_eu) 6.316  7.194 3.692 0.00808 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.456   Deviance explained = 57.9%
GCV = 8.0605e+11  Scale est. = 6.027e+11  n = 29
```

Summary 2

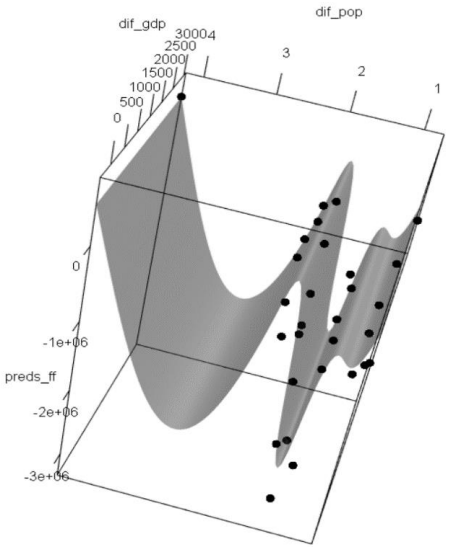

Figure 21

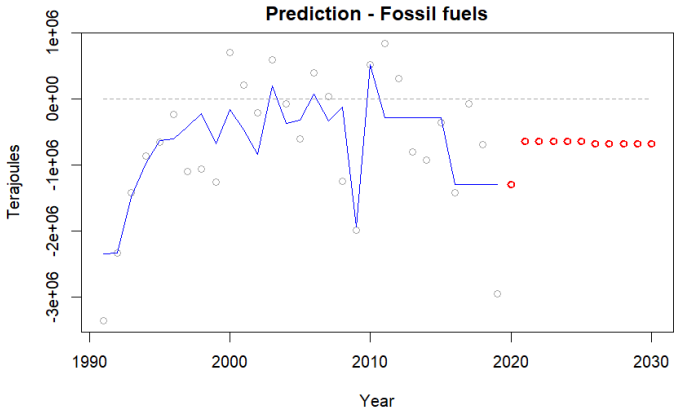

Figure 22

3. DIFF_EU_NATURAL_GAS ~ β_0 + DIFF_EU_POP + DIFF_EU_GDP : DIFF_EU_POP + ϵ

The obtained model considers a linear dependence from the population and the interaction of gdp and population. No gdp term, and no smoothing this time too.

We present the summary (Summary 3) of the model, along with the regression plane (Figure 23) and the temporal evolution plot (Figure 24).

```
Call:
lm(formula = dif_gas_eu ~ dif_pop_eu + dif_gdp_eu:dif_pop_eu,
    data = vec_gas.diff)

Residuals:
     Min       1Q   Median       3Q      Max
-1889489  -319899  -113618   692776  1103522

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1283937.4   555845.9   2.310  0.02909 *
dif_pop_eu          -996933.8   429256.2  -2.322  0.02830 *
dif_pop_eu:dif_gdp_eu     342.1      120.5   2.840  0.00865 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 793400 on 26 degrees of freedom
Multiple R-squared:  0.2375,    Adjusted R-squared:  0.1788
F-statistic: 4.048 on 2 and 26 DF,  p-value: 0.02948
```

Summary 3

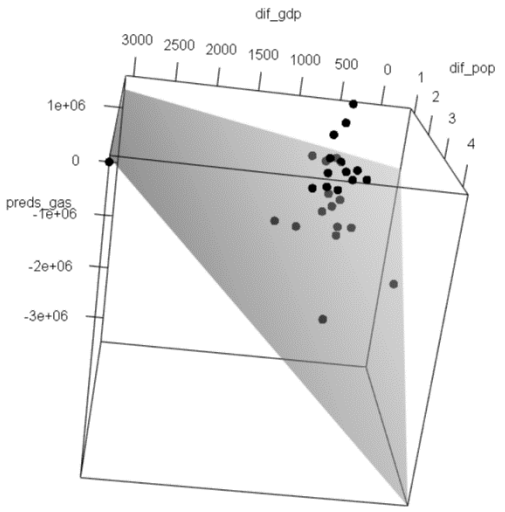

Figure 23

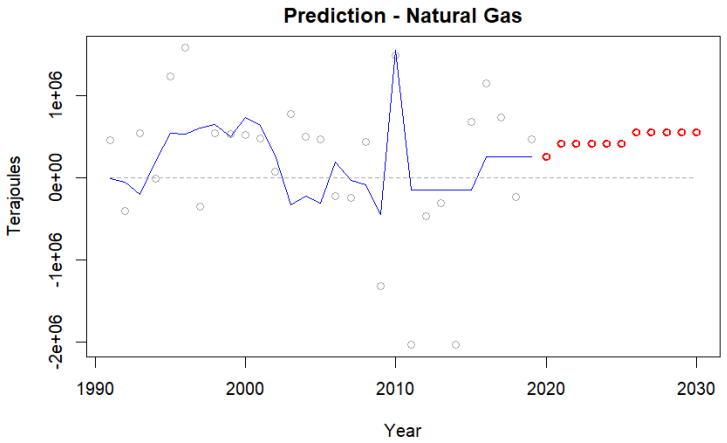

Figure 24

The last auxiliary model we fit is the one that finds a relationship between the renewable energy consumptions and the non-renewable ones, still considering year-to-year variations of both. Following the same procedure described above, we obtained the following model:

$$DIFF\_EU\_RENEWABLES \sim \beta_0 + f(DIFF\_EU\_OIL\&PETROLEUM) + f(DIFF\_EU\_FOSSIL\_FUELS) + f(DIFF\_EU\_FOSSIL\_FUELS * DIFF\_EU\_NATURAL\_GAS) + f(DIFF\_EU\_OIL\&PETROLEUM * DIFF\_EU\_NATURAL\_GAS) + f(DIFF\_EU\_OIL\&PETROLEUM * DIFF\_EU\_FOSSIL\_FUELS * DIFF\_EU\_NATURAL\_GAS) + \epsilon$$

The obtained Gam model considers all the original terms, except for the Natural Gas and the Oil&Petroleum - Fossil Fuels interaction, all smoothed with a cubic basis. We present the summary (Summary 4) of the model, along with the temporal evolution plot (Figure 25). We notice that the predictions are (strangely) slightly negative, but we will obtain a more reasonable result in the robustness section.

```
Family: gaussian
Link function: identity

Formula:
value_re ~ s(value_oil, bs = "cr") + s(value_ff, bs = "cr") +
    s(I(value_ff * value_ng), bs = "cr") + s(I(value_oil *
    value_ng), bs = "cr") + s(I(value_oil * value_ff *
    value_ng), bs = "cr")

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4154.5      319.3   13.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                                   edf Ref.df      F  p-value
s(value_oil)                     1.024  1.047 13.765 0.000379 ***
s(value_ff)                      7.920  8.589  7.604 1.69e-07 ***
s(I(value_ff * value_ng))        1.000  1.000 20.117 2.97e-05 ***
s(I(value_oil * value_ng))       6.691  7.335  3.717 0.002106 **
s(I(value_oil * value_ff * value_ng)) 4.939  5.519  7.354 9.98e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.63   Deviance explained = 72.5%
GCV = 1.1797e+07  Scale est. = 8.664e+06  n = 85
```

Summary 4

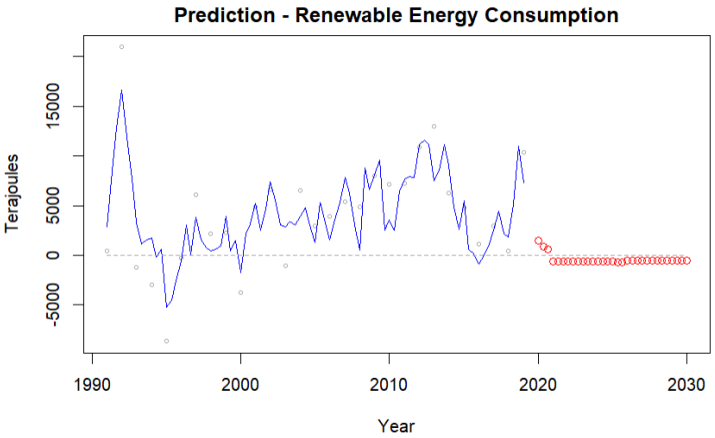

Figure 25

Having fit these auxiliary models, we could develop our final model for the prediction of future variations of Greenhouse gas emissions, until 2030.

The final model, obtained following the same procedure described above, is quite straightforward, in fact it is a linear model between the yearly variations of Greenhouse gas emissions and the yearly variations of renewable energies consumptions. We still exploited the Gam modelling to directly get an estimation of the degree of the regression curve, but eventually saw that a simple linear regression was enough. The variations of Ghg emissions are expressed as percentage, wrt to the 1990 value.

The final model

We present the summary of the model (Summary 5), along with the regression plot (Figure 26), the temporal evolution plot (Figure 27), the residuals Autocorrelation functions plots (Figures 28-29) and a plot of the conformal prediction (miscoverage level 0.05) on future values (Figure 30).

DIFF_EU_GHG_EMISSIONS ~ β_0 + DIFF_EU_RENEWABLES + ϵ

```
Call:
lm(formula = value_ghg ~ value_re, data = vec_ghg.diff)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5520 -1.0995  0.2794  0.8707  3.4347

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.412e-01  4.217e-01  -1.046   0.3047
value_re    -1.253e-04  6.057e-05  -2.069   0.0482 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.815 on 27 degrees of freedom
Multiple R-squared:  0.1369,    Adjusted R-squared:  0.1049
F-statistic: 4.281 on 1 and 27 DF,  p-value: 0.04824
```

Summary 5

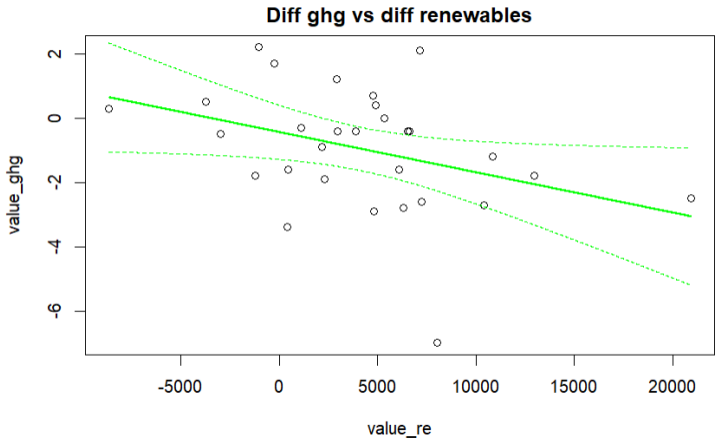

Figure 26

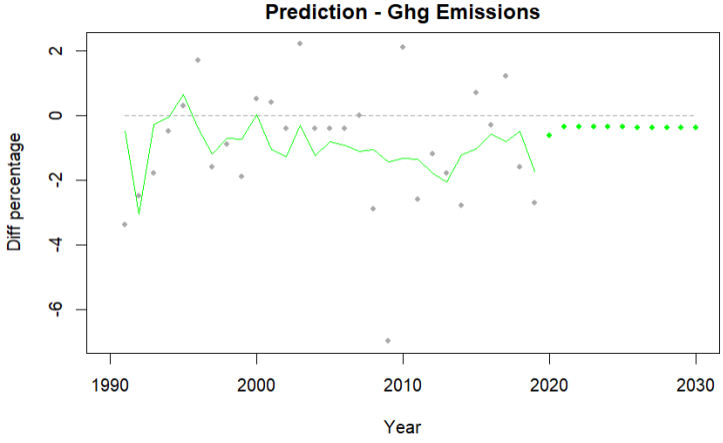

Figure 27

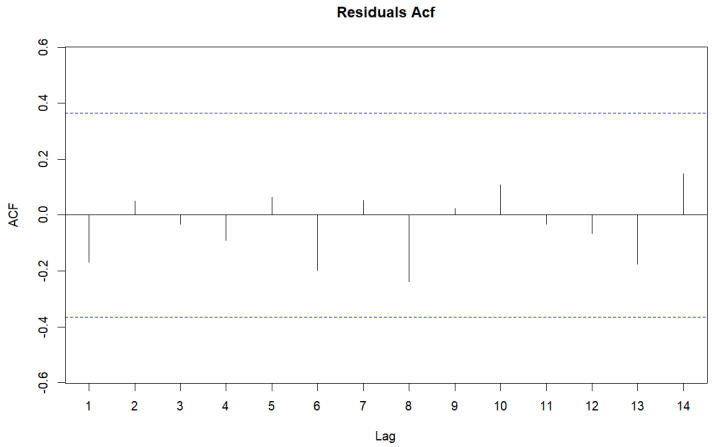

Figure 28

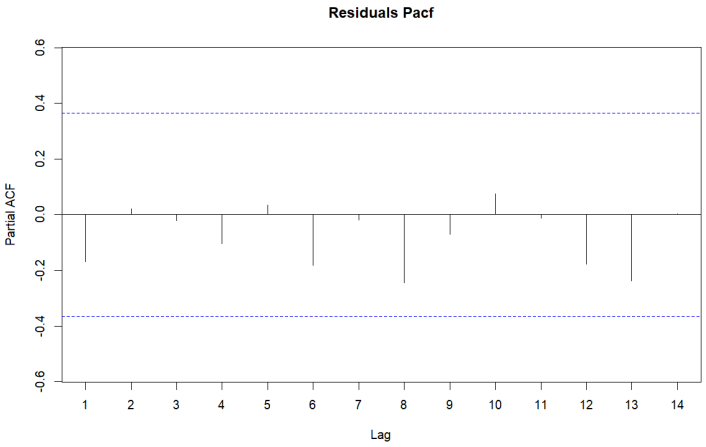

Figure 29

We notice how working with yearly variations permitted us to obtain uncorrelated residuals in the final model.

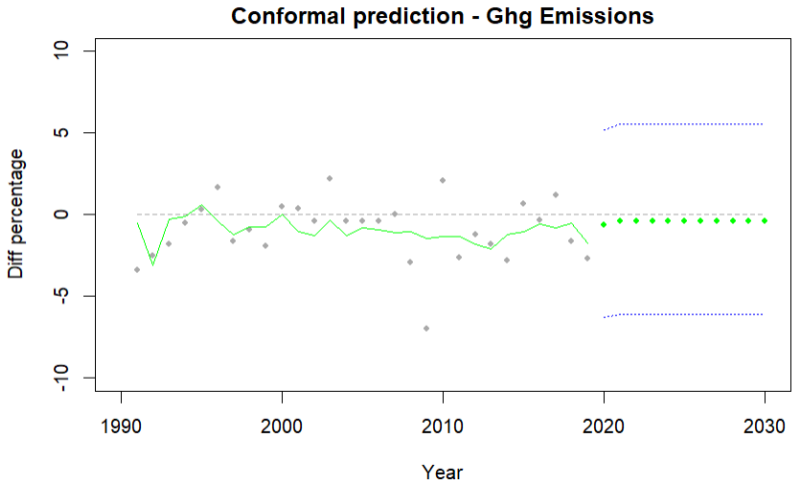

Figure 30

At this point we had all the tools to realize our much desired prediction:

How much will Greenhouse Gas emissions decrease by 2030? Is the Green Deal feasible?

Summing the yearly predicted variations we obtained a **predicted reduction of 32.3186 % (32.32%)**, as we can see in figure 31.

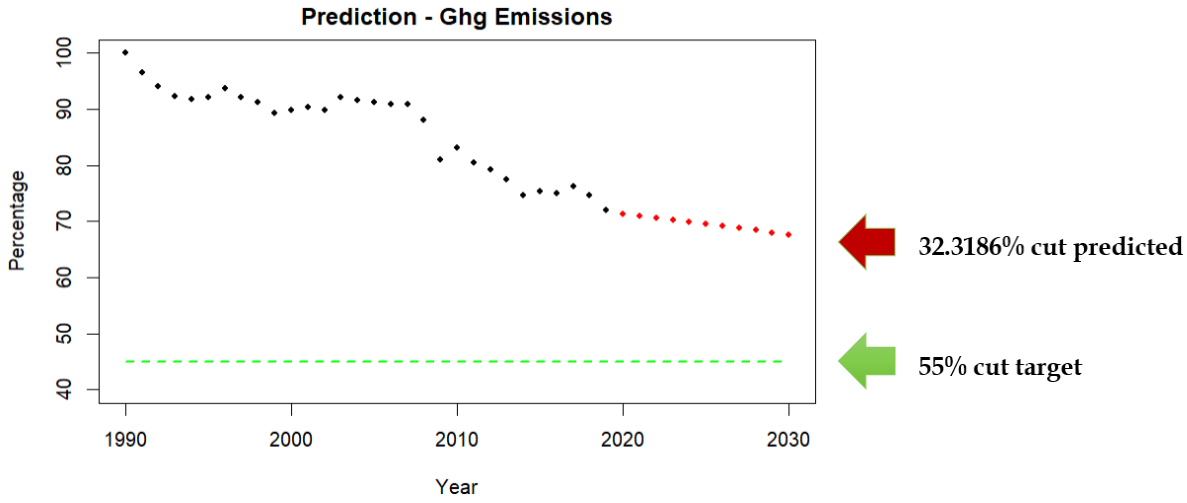

Figure 31

It is clear that **much more effort is needed** in order to comply with the Green Deal objective! The current rhythm is not enough.

ROBUSTNESS OF OUR RESULTS

In our project, we wanted to include a part about robustness methods applied to our datasets. Even though during the course's lectures all the methods regarding robust statistics were applied to multivariate datasets with no temporal correlation, some of the concepts can be used in our project to try to infer more information on our dataset.

Two specific methods were used in our code: Minimum Covariance Determinant and Robust regression.

During this section, when talking about energy datasets, we are using the year-to-year differences unless specified otherwise. It is also important to note that we used the value 0.75 for the alpha (α) parameter in MCD method.

Minimum Covariance Determinant

1. Outlying countries

We started with the objective of detecting outlier countries in our energy consumption datasets, to understand where our tests would have a "hard time".

This first approach consists in representing each country as an independent sample of p dimensions, where p corresponds to the value of each energy type for each year (N=Countries, p=energy sources).

This would mean we would obtain several country outliers for each year. A possible approach to join the information on outliers for all years would be counting how many times each country has been classified as an outlier. This way we would obtain a rank of the most outlying countries in our dataset.

Since for MCD to be applied we need a multivariate dataset, we have to combine at least two energy sources. So we try this method on the Non-Renewable energy sources (Fossil fuels, Oil & Petroleum and Natural Gas).

As an example of what we obtain for a single year see figure 32, where outliers are marked in red for Non-Renewable Energy Sources in 1991, with the relative Distance-Distance plot (figure 33). In this example the outliers detected are Netherlands (21), Poland(23), Romania (25) and Albania (33).

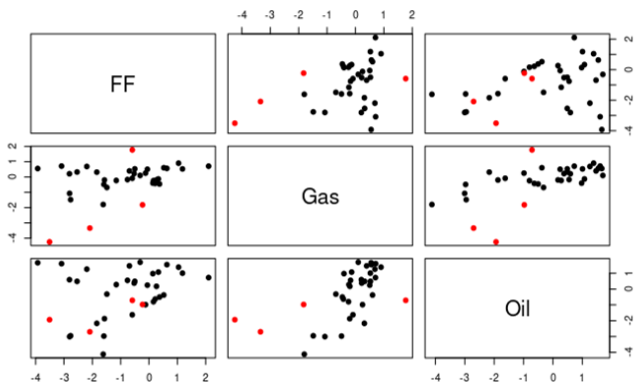

Figure 32

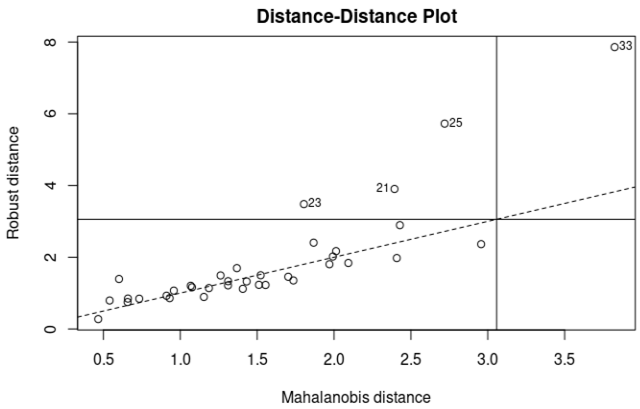

Figure 33

Using the ranking approach previously proposed we get the following table (Figure 34) of outlying countries from 1991 to 2019:

```
outliers
        Norway         Turkey North Macedonia        Finland     Luxembourg
             9              8                6              5              5
```

Figure 34

This section is quite useful from an analytic point of view, but not quite for our Core part of the project, since we use no specific country information in our final prediction models.

Our next subsection seemed us more useful for this purpose.

2. Outlying years

What if we try to detect which years have outlying properties in the same way as the previous subsection? This would mean transposing our data matrix, having years as rows. The interpretation in this case would be that our dataset is composed of 29 independent variables (from 1991 to 2019).

Regarding the representation of each of these 29 variables, we would use the information on the energy totals of the EU (N=years, p=energy sources). This would be interpreted as years for which we had an anomalous growth (or decrease) of total energy consumption with regard to the rest of the years.

A problem arised from this last proposal: should we include in the year axis the predicted values for future values? We tried both options, but we should take into account that NOT including the future values would mean we could try to extract the outliers from our original dataset and refit our models from a "new" dataset.

We first tried to detect outliers including the future predictions. As in the previous section, we used the combination of the three Non-Renewable energy sources' consumption levels.

We obtained the years 1991, 1992, 2009, 2010, 2011, 2014 and 2019 as outliers, with the following plots (Figure 35):

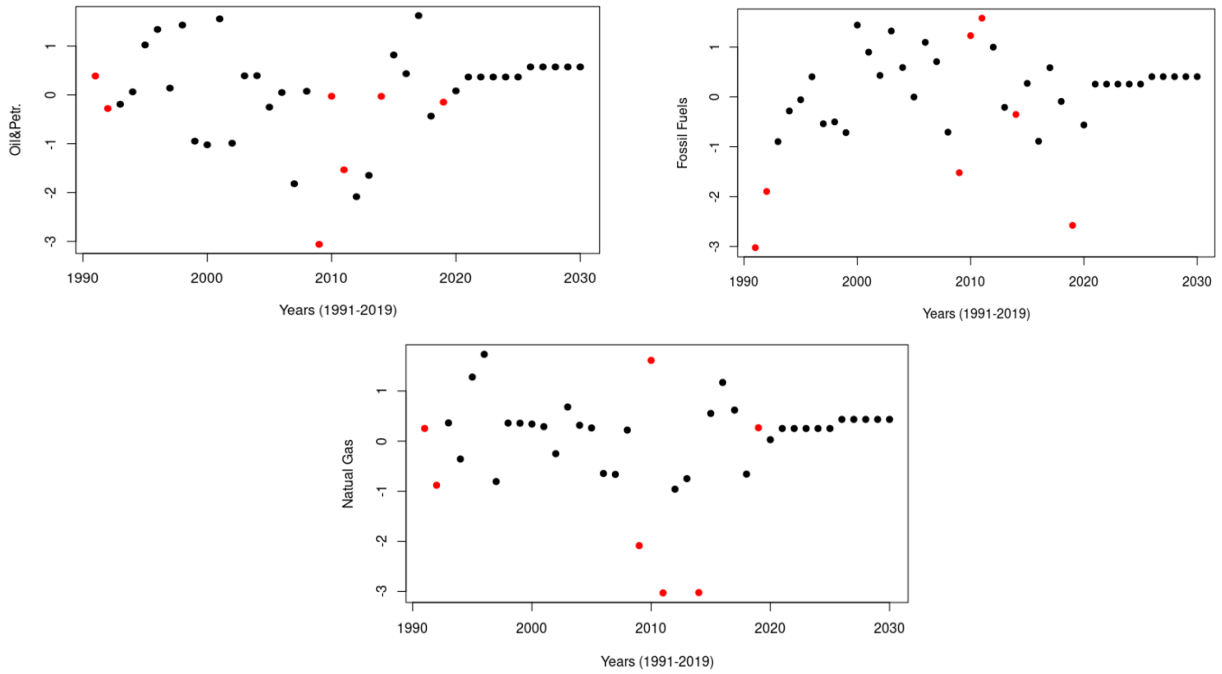

Figure 35

Let us try to use the second method proposed (i.e. MCD on true values without predicted ones). On the dataset composed of the years 1991-2019 we obtained the outliers 2011 and 2014 (Figure 36):

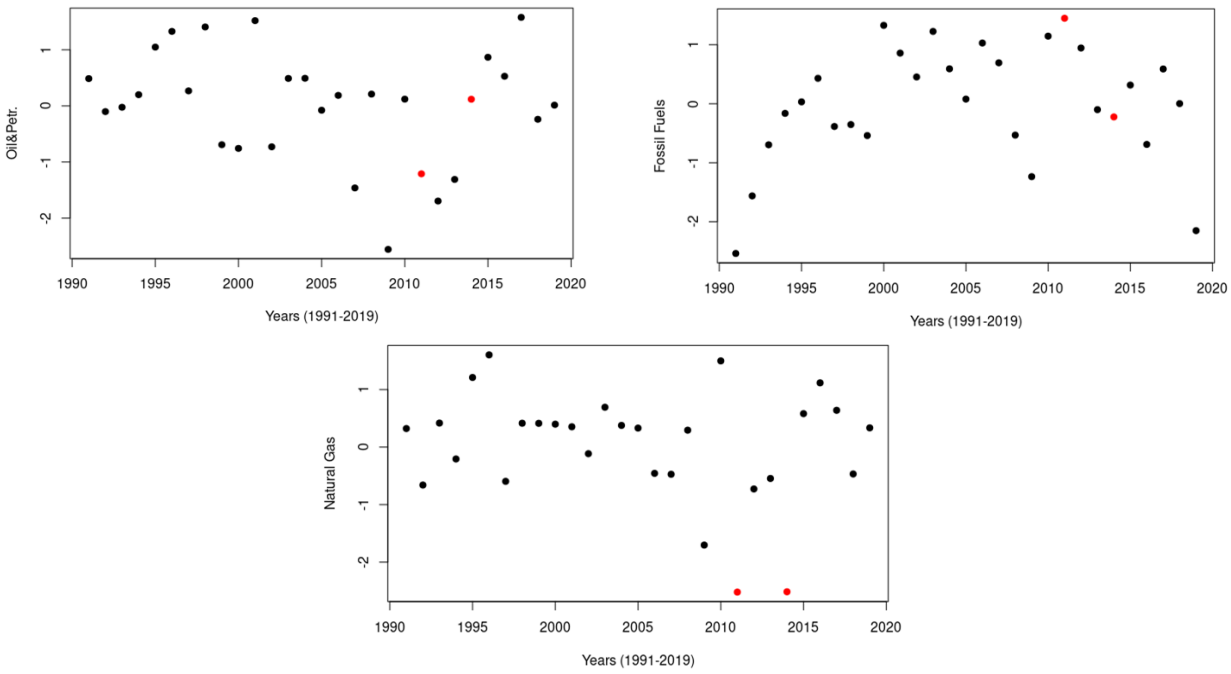

Figure 36

Inspecting the individual graphs and the distance-distance plots for these last two datasets, we can assume that it is better to choose the outliers detected in the current dataset since the 2011 and 2014 yearly NRE consumptions are much clearer outliers than the extra ones detected when including future predictions.

We can infer that the extra outliers are selected as such because of the extra-centrality that the predictions of Non-Renewable consumptions add to the dataset. It is clear that a swamping effect is taking place because some variabilities are amplified and falsely tagged as outlying.

Let us continue by refitting the Renewables ~ Non-renewables model to obtain Renewables predictions. After a brief model selection, we obtained a GAM model formulated by:

$$\text{DIFF\_EU\_RENEWABLES} \sim \beta_0 + f(\text{DIFF\_EU\_FOSSIL\_FUELS}) + f(\text{DIFF\_EU\_NATURAL\_GAS}) + f(\text{DIFF\_EU\_FOSSIL\_FUELS} * \text{DIFF\_EU\_NATURAL\_GAS}) + f(\text{DIFF\_EU\_OIL\&PETROLEUM} * \text{DIFF\_EU\_FOSSIL\_FUELS} * \text{DIFF\_EU\_NATURAL\_GAS}) + \epsilon$$

where oil and oil-fossil fuels interaction are discarded. We obtain the following positive predictions (Figure 37), which are a bit more reasonable than the ones obtained with the original model:

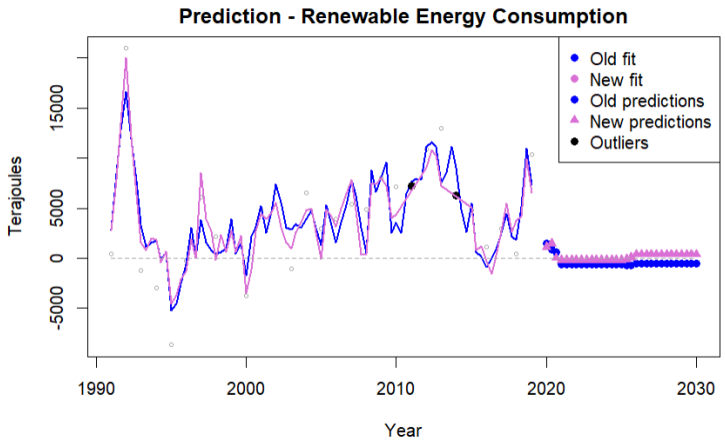

Figure 37

Next we retrained again the Greenhouse Gases regression model with a single regressor (Renewable Energy) with the newly found dataset without the 2011 and 2014 information.

The GAM machinery still suggest us to use a linear model, however unfortunately now the model is not significant at 5% (p-value = 0. 071). Still, the purpose of this model is comparison, so we stick with it.

Below we plot the two regression lines (Figure 38), the one obtained with our original model, together with the one obtained with the new model.

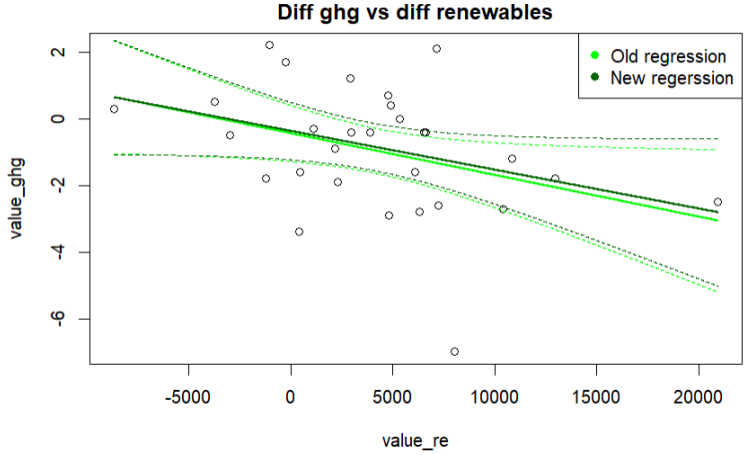

Figure 38

We also plotted the comparison between the old and the new predictions obtained for future Greenhouse gas emissions (Figure 39).

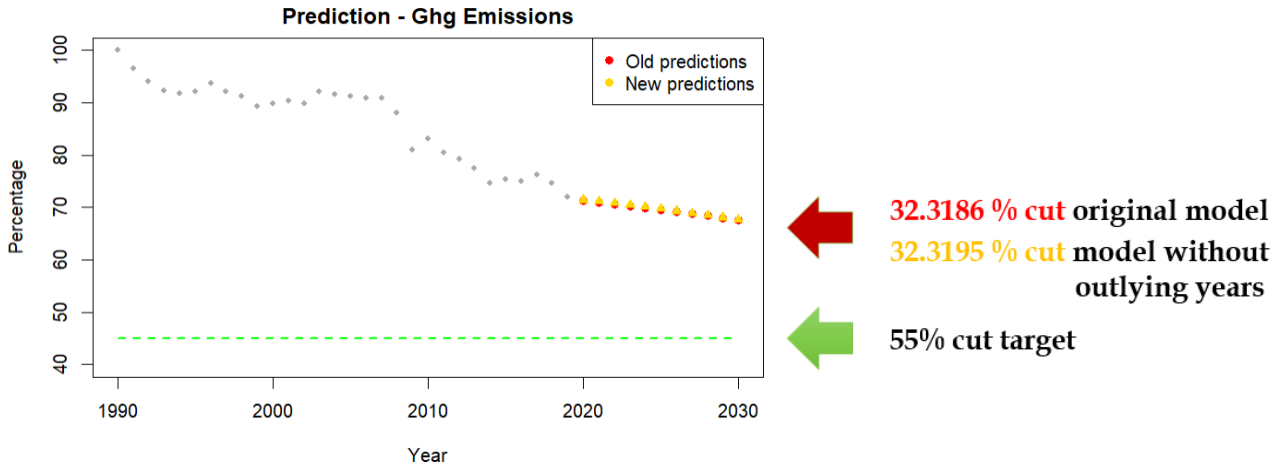

Figure 39

Robust Regression

Another approach was to use Robust Linear models on our final predictor i.e. Greenhouse Gases ~ Renewables. This made a lot of sense since our original model is a linear one. We expected for these models to act not too worse than the original linear model. This means we would try to make inference by using this methods on our original data.

We applied the Least Median Squares (LMS) and Least Trimmed Squares (LTS) models, covered during the course. For the former, we used the *lmsreg(.)* function, while for the latter we used the *lmrob(.)* function. Contrary to the course's labs, we decided to use *lmrob(.)* instead of *ltsReg(.)* following the robustbase package documentation's advice (*see References*): "We strongly recommend using *lmrob(.)* instead of *ltsReg(.)* […]".

When fitting both models and comparing it with the original linear model, we saw a growth in sum of square of residuals and a decline on the adjusted R-squared, even though for the LTS model the differences are very little.

Let us take a look at the plots for the regression of each of the models w.r.t. the original one (Figure 40).

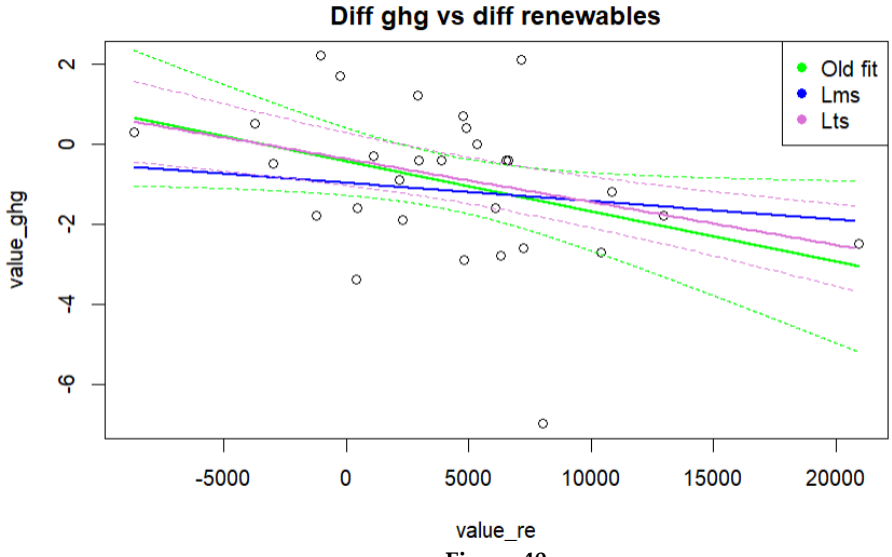

Figure 40

The LTS model seems to be very similar by looking at the graph above.

Let us take a look at the LTS predictions for 2020-2030, compared with the original ones (Figure 41):

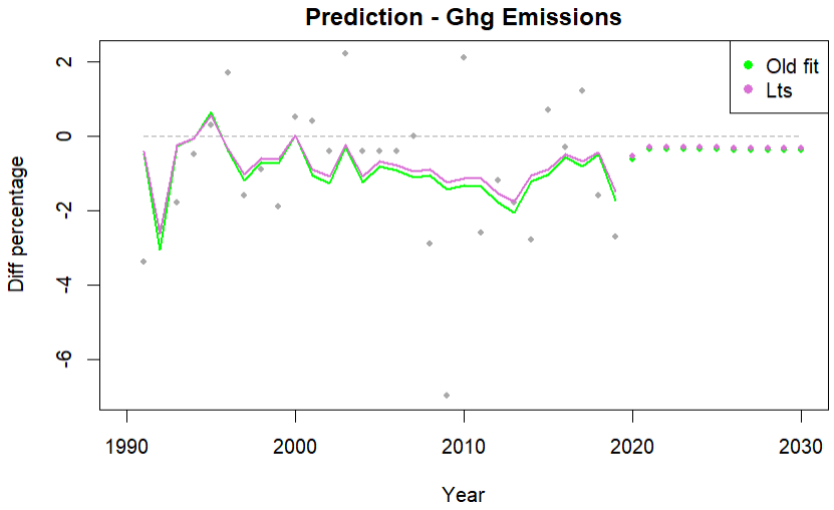

Figure 41

As before, we also plotted the comparison between the old and the new predictions obtained for future Greenhouse gas emissions, with this other approach (Figure 42).

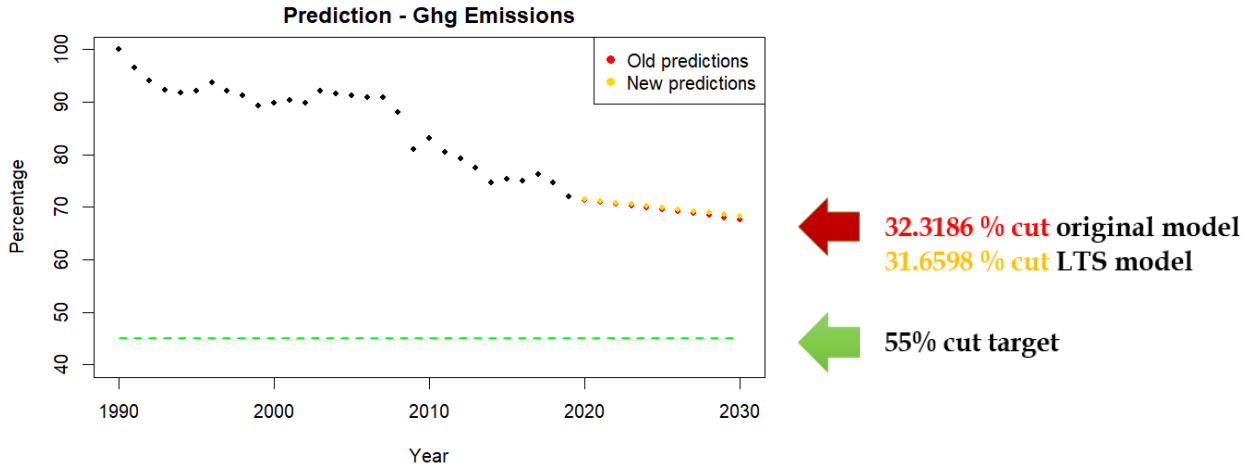

Figure 42

The predictions look again very similar to the ones obtained with the original model.

Robustness Conclusions

We find that the predictions obtained by using the LTS estimator are almost the same, even a little more conservative since the decrease is a little less tilted than the LM one.

From this section's LTS model, which tries to robustify a linear model, we find that, since only one outlier (index 19) is trimmed away, there is almost no difference between our original LM model and this new one.

This could be understood as the possibility for our project to choose the LTS model as a final predictor for the GHG emissions. A balanced approach between robustness and accuracy should be taken into account. Both models seem valid enough.

Even the predictions made by quitting the outlying years from the non-renewable sources datasets are still very similar to the one obtained with the original data, so all this evidence makes us confidently say that our models and predictions are quite robust, since they are not influenced that much by the presence of outlying observations.

CONCLUSION

Global warming is happening and it is strongly tangible. It is not a fantasy of some crazy scientist or some conspiracy theorist. It is a sad reality that will have heavy consequences on everybody's life, if we don't act immediately. It is everybody's responsibility, from the major multinational companies to the simple citizens, to become aware of the problem and to take the appropriate countermeasures.

With the recent European agreements, our leaders showed the right intention of putting this problem in the spotlight, to guarantee a sustainable future for our planet. This, however, doesn't seem to be enough.

Our work, that highlighted the non-feasibility of the Green Deal target of a 55% cut in Greenhouse gas emissions by 2030 respect to 1990, wants to be yet another raised alarm, to be added to the already long list of scientific researches on this subject.

We are delighted to have had the possibility to apply the Nonparametric Statistics course tools to better grasp this trending and so important topic.

Special thanks to all the professors for the opportunity to get our hands dirty, analysing real data, in order to develop a project with concrete and useful conclusions.

REFERENCES

Datasets taken from EU's Eurostat portal: https://ec.europa.eu/eurostat/web/energy/data/database

Complete energy balances (gross available energy) dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_bal_c&lang=en

Consumption of solid fossil fuels (inland consumption) dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_cb_sff&lang=en

Consumption of oil and petroleum products (gross inland deliveries) dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_cb_oil&lang=en

Consumption of natural gas (inland consumption) dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_cb_gas&lang=en

Consumption of renewables and wastes (inland consumption) dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_cb_rw&lang=en

Renewable energy sources percentage over total energy available dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_ind_ren&lang=en

Consumption of electricity (gross electricity production) dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_cb_e&lang=en

Use of renewables for electricity dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_ind_ured&lang=en

Electricity production capacities by main fuel groups dataset:
https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_inf_epc&lang=en

Greenhouse Gas emissions (percentage wrt 1990 value) dataset:
https://ec.europa.eu/eurostat/databrowser/view/SDG_13_10/default/table

Population EU 1990-2021:
https://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD?end=2020&locations=EU&start=1990

GDP EU:
https://data.worldbank.org/indicator/SP.POP.TOTL?locations=EU

Bureau of Labour Statistics CPI inflation calculator:
https://data.bls.gov/cgi-bin/cpicalc.pl

SSPs:
Keywan Riahi, Detlef P. van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C. O'Neill, Shinichiro Fujimori, Nico Bauer, Katherine Calvin, Rob Dellink, Oliver Fricko, Wolfgang Lutz, Alexander Popp, Jesus Crespo Cuaresma, Samir KC, Marian Leimbach, Leiwen Jiang, Tom Kram, Shilpa Rao, Johannes Emmerling, Kristie Ebi, Tomoko Hasegawa, Petr Havlík, Florian Humpenöder, Lara Aleluia Da Silva, Steve Smith, Elke Stehfest, Valentina Bosetti, Jiyong Eom, David Gernaat, Toshihiko Masui, Joeri Rogelj, Jessica Strefler, Laurent Drouet, Volker Krey, Gunnar Luderer, Mathijs Harmsen, Kiyoshi Takahashi, Lavinia Baumstark, Jonathan C. Doelman, Mikiko Kainuma, Zbigniew Klimont, Giacomo Marangoni, Hermann Lotze-Campen, Michael Obersteiner, Andrzej Tabeau, Massimo Tavoni.
The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview, Global Environmental Change, Volume 42, Pages 153-168, 2017, ISSN 0959-3780, DOI:110.1016/j.gloenvcha.2016.05.009

Oliver Fricko, Petr Havlik, Joeri Rogelj, Zbigniew Klimont, Mykola Gusti, Nils Johnson, Peter Kolp, Manfred Strubegger, Hugo Valin, Markus Amann, Tatiana Ermolieva, Nicklas Forsell, Mario Herrero, Chris Heyes, Georg Kindermann, Volker Krey, David L. McCollum, Michael Obersteiner, Shonali Pachauri, Shilpa Rao, Erwin Schmid, Wolfgang Schoepp, Keywan Riahi,
The marker quantification of the Shared Socioeconomic Pathway 2: A middle-of-the-road scenario for the 21st century, Global Environmental Change, Volume 42, 2017, Pages 251-267, ISSN 0959-3780, DOI:10.1016/j.gloenvcha.2016.06.004

LtsReg for robust regression:
https://rdrr.io/cran/robustbase/man/ltsReg.html