LIÈGE université

FACULTÉ DES SCIENCES APPLIQUÉES
Ecole d'ingénieurs et d'informaticiens

PROJ0016-1
BIG DATA PROJECT

# Will EU achieves its goals of $CO_2$ reduction by 2030?

MILESTONE 3

VIESLET Thomas - s153205
GIRINEZA Guy - s144377
DELVOYE Benjamin - s154317

*Master 1 Data science Engineering*
*Faculty of Applied Sciences*

Academic years 2019-2020

# 1 Variables selection

## 1.1 Why fossil energy consumption and not primary energy consumption ?

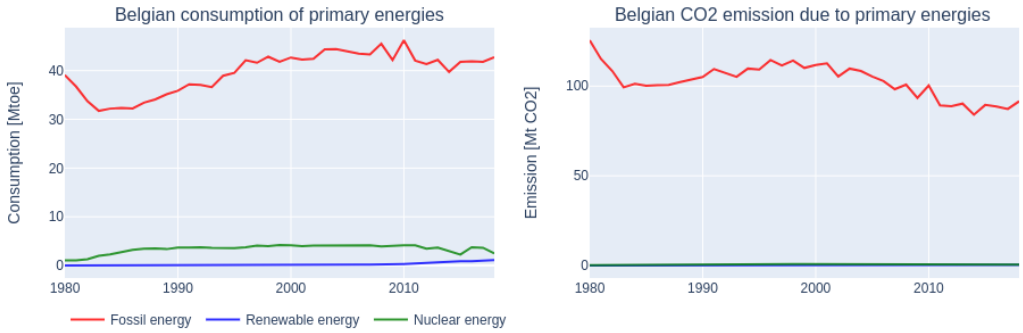


Figure 1

It is observable on figure 1 that since 1996, $CO_2$ emissions due to primary energies[1] have decreased in Belgium while since 1983, the consumption of those primary energies have increased. Among these primary energies, some can be considered as carbon free. CO2 emissions due to the consumption of renewable and nuclear energy are negligible compared to emissions due to fossil fuels. It can be observed on figure 2 where the quantity of CO2 emitted by unit of each type of primary energy consumed is displayed. This is why we didn't include CO2 emissions of renewable and nuclear energies.

[1] Fossil energies (coal, gas, oil), nuclear energy and renewable energies (solar, wind, hydraulic)

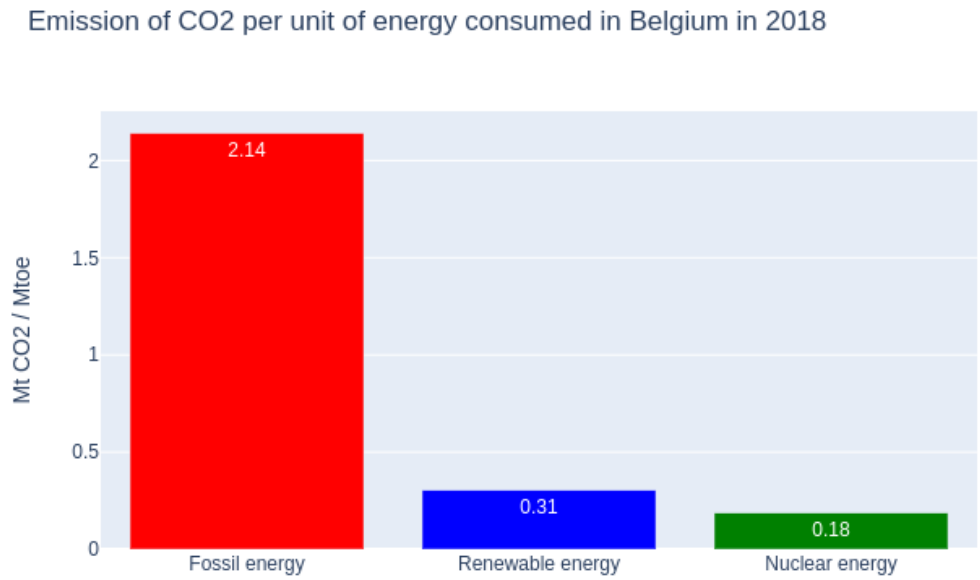


Figure 2

We chose to build our models on top of those explanatory variables:

- Population.
- Gross domestic product per capita.
- Carbon intensity of fossil energy:
  Number of mega tonne $CO_2$ over mega tonne oil equivalent of fossil energy.
- Primary intensity:
  Number of mega tonne $CO_2$ over gross domestic product per capita.
- Nuclear consumption:
  Mega tonne of oil equivalent.
- Renewable consumption:
  Mega tonne of oil equivalent.

# 2 Linear regression

The first model we are going to use to forecast the consumption of fossil energy is the famous linear regression. We used the *LinearRegression* function of *scikit-learn*. This version uses ordinary least square (the maximum likelihood estimator) in order to estimate the coefficients of the unknown parameters in a linear regression model. The goal is to minimise the sum of squared difference of distances between the data points of the observed dependant variables and those predicted by the regression line.

Before using a linear regression that uses ordinary least square, it is important to check if our data is appropriate for such a model. In order to verify it, we use the Gauss-Markov theorem. The Gauss-Markov theorem is telling us that ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate (BLUE) possible, given that all Gauss-Markov assumptions are fulfilled. Those assumptions are the following :

- Linear relationships
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity
- The error term is normally distributed

## 2.1 Linear relationships

This model needs linear relationship between the independent variables and the dependant variable. You can verify this assumption by plotting a scatter plot for each independent variable in function of the dependant variable.

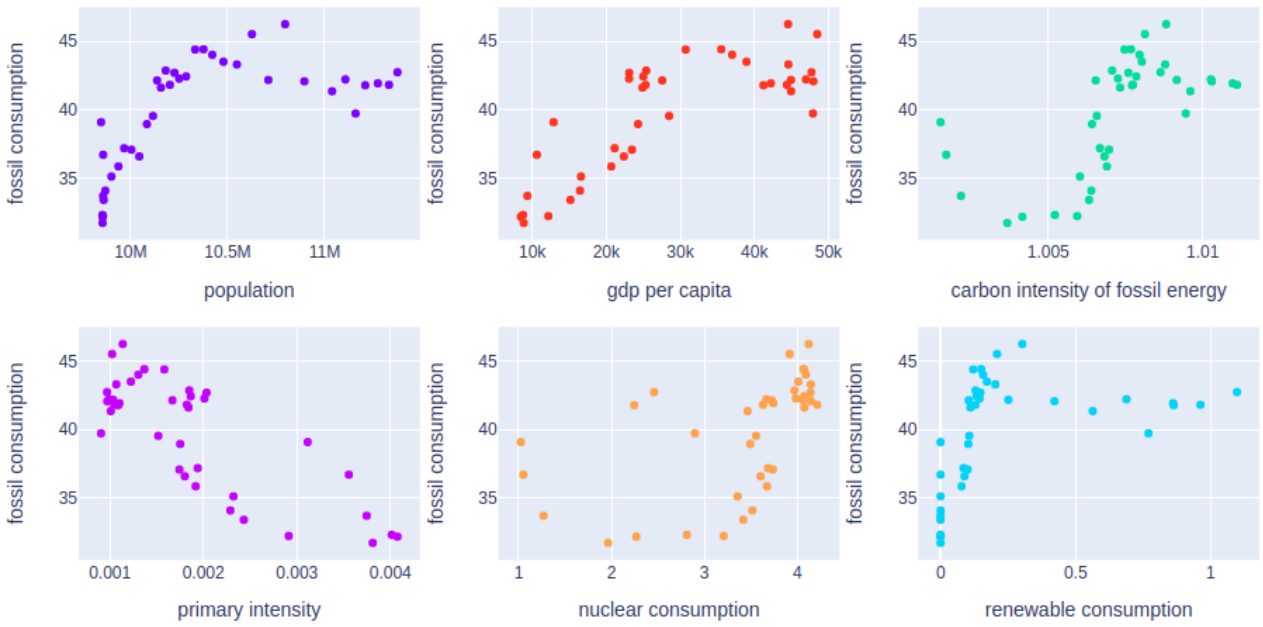


Figure 3: Scatter plots of fossil consumption in terms of the independent variables

Variables such as Population, primary intensity, gdp per capita and carbon intensity of fossil energy shows more or less signs of linear relation with fossil consumption contrary to nuclear and renewable consumption variables.

## 2.2 No or little multicollinearity

Multicollinearity occurs when there exists near-linear relationships among independent variables. In other words, when independent variables are too highly correlated with each other. In a case of multicollinearity, least square estimates are unbiased, but their variance are large and so the values of those estimated coefficients might be far from the true value. In order to detect multicollinearity, you can use pairwise scatter plots between independent variables or correlation matrix.

| | population | gdp per capita | Carbon intensity of fossil energy | primary intensity | nuclear consumption | renewable consumption |
|---|---|---|---|---|---|---|
| **population** | 1 | 0.91 | 0.8 | -0.76 | 0.23 | 0.94 |
| **gdp per capita** | 0.91 | 1 | 0.85 | -0.91 | 0.47 | 0.75 |
| **Carbon intensity of fossil energy** | 0.8 | 0.85 | 1 | -0.87 | 0.73 | 0.68 |
| **primary intensity** | -0.76 | -0.91 | -0.87 | 1 | -0.65 | -0.62 |
| **nuclear consumption** | 0.23 | 0.47 | 0.73 | -0.65 | 1 | 0.037 |
| **renewable consumption** | 0.94 | 0.75 | 0.68 | -0.62 | 0.037 | 1 |

Figure 4: Correlation matrix of independent variables

Unsurprisingly, some independent variables are highly correlated with each other. We will have to pay close attention to the values of the coefficients that we will obtain with linear regression, both in terms of their scales and their signs.

## 2.3 No auto-correlation

Auto-correlation occurs when the residuals[2] are not independent from each other. So when the observation of an error term can predict the next observation. In order to detect auto-correlation of each independent variables, one need to plot the scatter plot of each of these variables in terms of the residuals. Then, we check if the data is well distributed without any specific pattern.

[2]Difference between the observed values and the predicted values

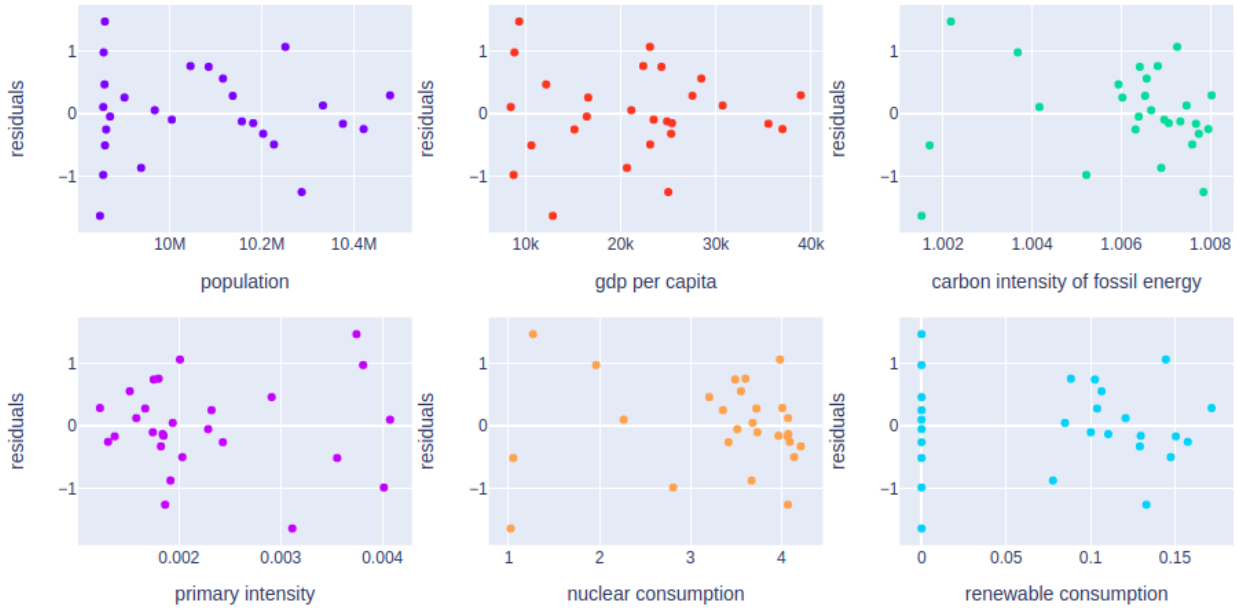


Figure 5: Scatter plots of residuals in function of the independent variables

On figure 5 we can observe a cone shape being drawn with the increase of multiple variables such as population, gdp per capita, primary intensity or carbon intensity of fossil energy. We can easily conclude that most of our independent variables are auto-correlated which is not surprising when working with time series. The coefficients obtained with OLS will not be BLUE and will not be reliable enough.

## 2.4 Homoscedasticity

This condition means that the residuals should be consistent for all observations. In other words, the variance of the residual do not vary from one observation to another. Since we're working with time series, one can check if this assumption is violated by looking at the plot of the residuals with respect to time.

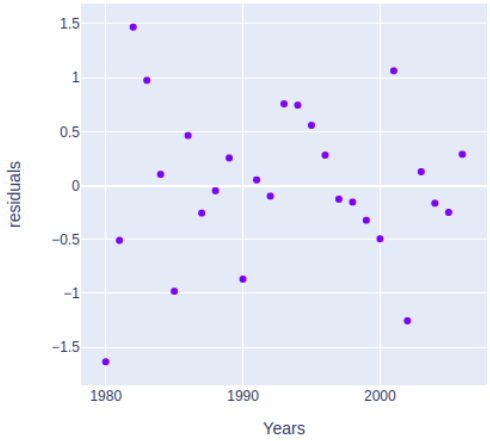


Figure 6: Scatter plots of residuals in function of time

The data looks pretty spread out around the x-axis, which makes us conclude that homoscedasticity assumption is violated by our data. When the database is large, this hypothesis is generally respected. In our case, we only work with 39 data.

## 2.5 The error term is normally distributed

In reality, linear regression used with OLS do not need the error term to be normally distributed in order to compute the unknown parameters. However, this assumptions is necessary if you want to prove that the least squares estimators are BLUE.

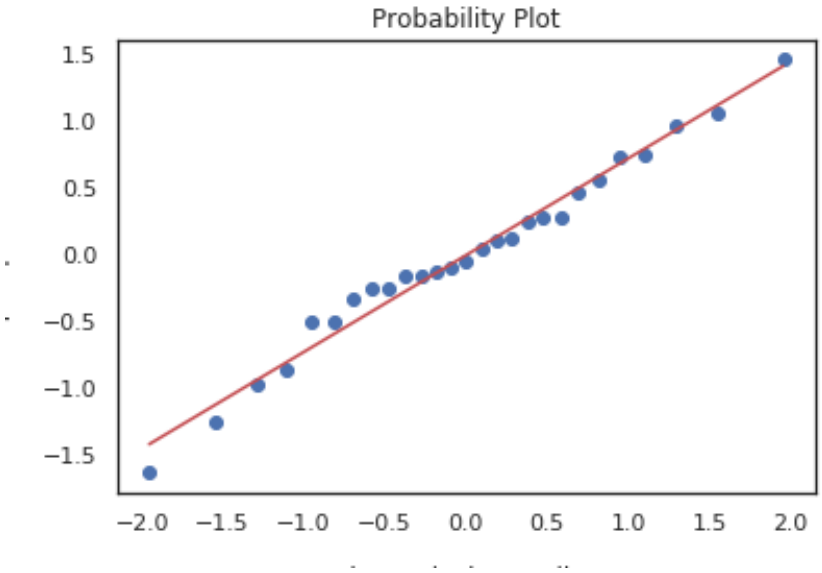


Figure 7: Probability plot

This assumption can be considered as fulfilled.

In conclusion, the estimators of the OLS are not BLUE and should be used with precaution. However, some of the unexpected assumptions might be fixed. By example, multicollinearity can be managed by a ridge regression or and most of the issues can be man selection technique.

## 2.6 Fitted model

As training set, we use value of the dataset from 1980 to 2006 and for the validation set, one use values from 2006 to 2018. The result of the linear regression is the value of the intercept $\beta_0$ and the value of the coefficient $\beta_i$ of each independent variables.

- $\beta_0 =$ 6306.06
- $\beta_1 =$ 2.27e-05
- $\beta_2 =$ -1.10e-04
- $\beta_3 =$ -6.48e+03
- $\beta_4 =$ -1.94e+03
- $\beta_5 =$ 1.06e+01
- $\beta_6 =$ 9.648

Let's analyse the scale and the sign of these coefficients. In order to do it looking at figure 3 and at the scatter plots of the independent variable in function of time represented by figure 17[3].

But first, one need to remind what this coefficient represents. A coefficient of a linear regression gives you the evolution of the dependant variable for one unity of the corresponding independent variable. The sign of this coefficient gives the direction of the relationship between an independent variable and the dependent variable.
$\beta_1$ et $\beta_4$, the signs of the coefficients associated respectively to population and primary energy intensity variables presents coherent behaviour. It is clearly harder to say for nuclear ($\beta_5$) and renewable ($\beta_6$) energy consumption by looking to their scatter plots with respect to the fossil consumption but more understandable when observing figure 17. And finally, $\beta_2$ et $\beta_3$, the signs of the coefficients associated respectively to gdp per capita and carbon intensity of fossil energy variables are clearly abnormal. Both of them should be positive. It is another clear sign of multicollinearity !

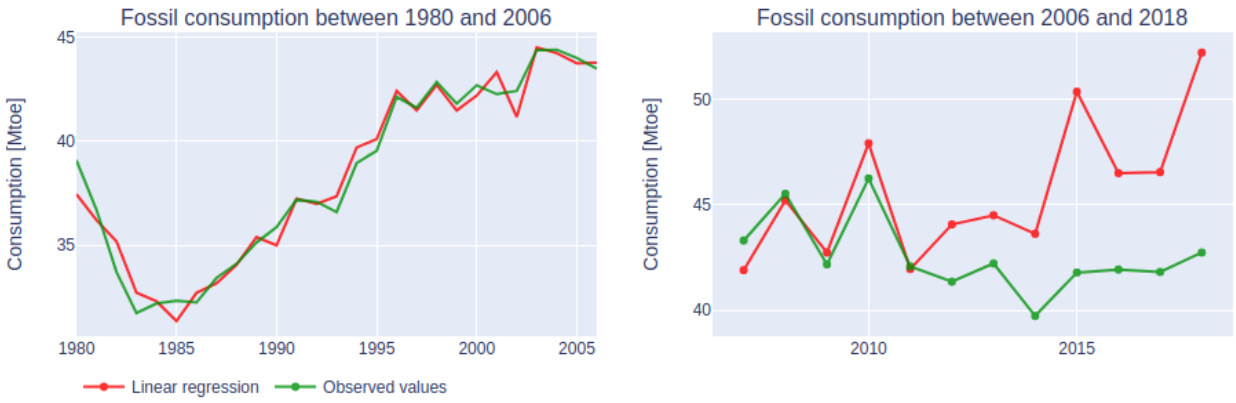


Figure 8: Results of the linear regression model

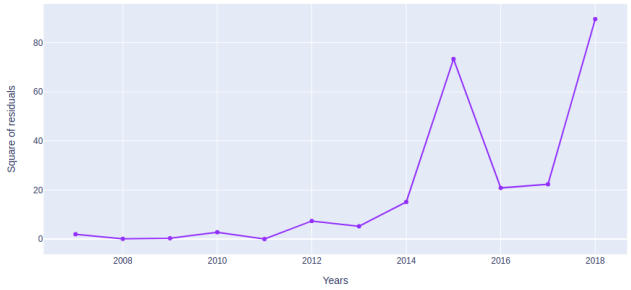


Figure 9: Square residuals in function of time

Looking at figure 8, one can observe that this model fit well the the data. It can be due to overfitting. More interesting is the results this model gives us for the validation set. Indeed, the five first observed data are well predicted by the OLS estimators while from 2012 to 2018, the square residuals grows bigger and bigger. From those results, we can conclude that this model can be 'trusted' for a forecasting of five years ahead. However, you have to take a lot of precautions due the violation of some Gauss-Markov assumptions, especially the one on multicollinearity.

[3]See annex

# 3 No auto-correlation

# 4 Bayesian Ridge Linear regression

## 4.1 Motivations

Bayesian linear regression is the Bayesian view of the linear regression while the classical least square error regression is based on a frequentist approach. Indeed, classical linear regression consider the sample points as point estimates and the Bayesian regression consider that it is the value of a random variable drawn from a probability distribution (here, as often, from a Gaussian distribution). The parameters of the problems as the precision of the coefficients can also be considered as drawn from probability distribution.

There are a few reasons why we tried this method in our present project. We had previously decided that we would focus on simple linear methods for the time being and had to find another linear method then the classical one. Bayesian regression seemed specially suited for the following reasons:

### Prior knowledge

Theoretically, the Bayesian regression is able to make use of some prior knowledge about the data. We thought that it would be a good opportunity to increase the importance of the closest years data. However, the *sklearn* method makes use of gamma functions for the probability distributions of its parameters $\lambda$ and $\alpha$ (which role stayed unclear to us). This effect made it to difficult to take advantages of this property.

### Suited configuration

As the used method is Bayesian **Ridge** regression, this method gets also some properties of the Ridge regression (Ridge is penalisation that can be applied to a variety of models) as the fact that it is especially suited for high number of dimensions (which can be more or less the case here compared to the size of our dataset).

Furthermore, the Bayesian regression itself is said to be efficient in the case of limited data and more suited to ill-posed problem than least square error.

### Uncertainty

Finally, the main reason why we choose to lean on this method is because of the possibility to compute and predict a density of probability and so, to have a very good idea of the uncertainty of our model (as it has been asked of us in the previous milestone). Indeed, this is not something possible with the classical linear regression as it only computes point estimates.

## 4.2 Results

### Set-up

The whole model is trained on the first 29 years and tested on the last ten years. The following results are present on this period. We tried a variety of parameters configurations. Obviously, the default configurations was included but we obtained diverging behaviour which was ending pretty far from true values after 10 years. The most relevant configuration is with a single parameter change. The parameter *lambda_1* (shape parameter for the Gamma distribution prior over the lambda parameter which is the prior which is the precision of the prior Gaussian

distribution of the coefficients) is set to 550 000 which is far from default value. The only explanation we have for this result is intuitive.

As an increase of the shape parameter of a gamma function flattens the distribution and shift it to the left, the latter will come closer to a uniform probability distribution. This has for effect that the lambda (precision) parameter prior probability distribution is close to a uniform distribution and so, permits higher values of variance. As we have very different variables with very different values and variances, it can be considered good practise to have high-variance and no too-restrictive distribution of it (specially near 0).

### Correlation between fitted and true values

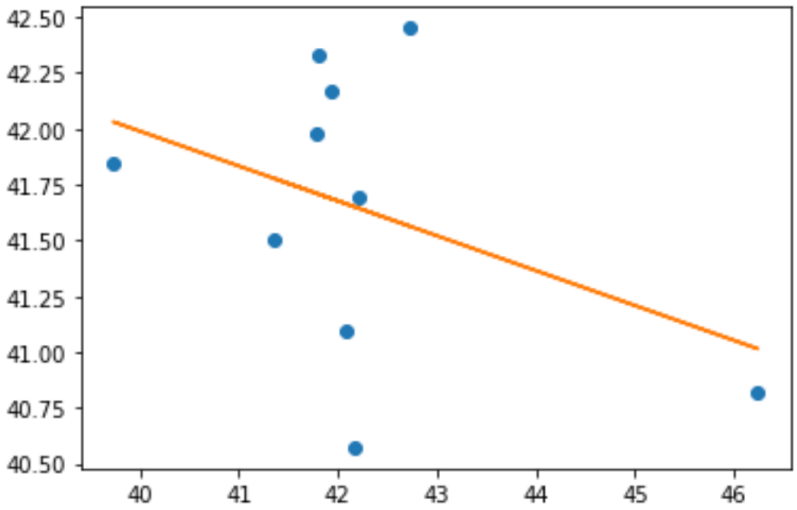


Figure 10: Correlation between fitted and real values plot

At first sight, this resulted graph seems to show a bad model as the correlation is supposed to be linear (near the y = x line). However, at first, it must be noticed that the 2 axes are not on the same scale and so, the highest group of points is not actually so far from the (y=x line). Secondly, the computed line is actually computed with a simple least square linear regression and is impacted extraordinary values.

If we consider that the scope of our project is not to prevent extraordinary values due to extraordinary events (which would be very difficult ten years ahead), this result may still be rather interesting.

**Actual predicted values for fossil consumption**

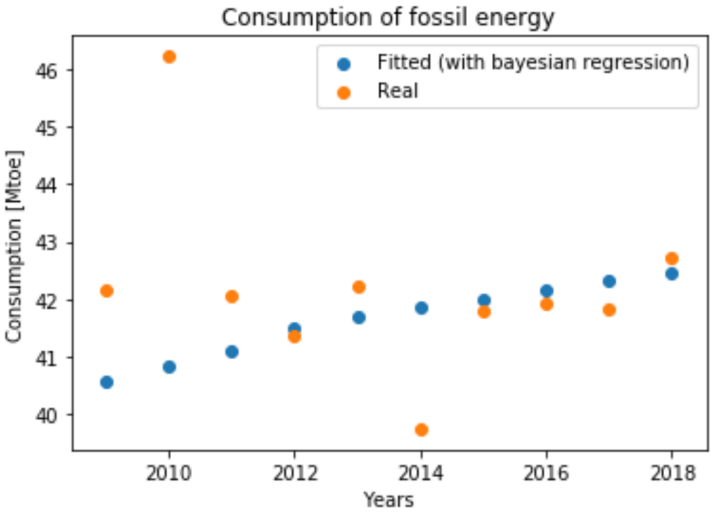


Figure 11: Fitted and real values plot

While looking at the above plot, we can infer a few thing. First, the extreme values are really bad fitted which was explained and understood from the previous plot. Secondly, the first fitted values are not so close to the real values. This is probably our biggest problem as the fitted value should be getting further from the real values with time. Finally, our fitted values are very close to the real value for the final year 2018 which is good news as our goal is to predict the consumption ten years from now.

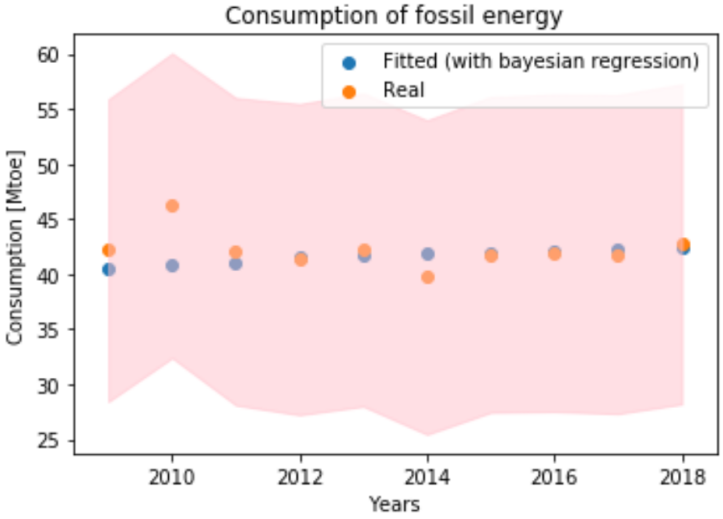


Figure 12: Fitted and real values plot with uncertainty

As we can see on the uncertainty plot, our model is very uncertain which does make sense considering the complexity of the problem. However, the uncertainty should intuitively be smaller near 2010 and increase with time which is not the case. It is another point to look into.

### 4.3 Remarks

The obtained results seems to show a relative stiffness of the model and furthermore, the computed coefficients are very small which makes us doubtful considering this model. However, this configuration ensures that we avoid overfitting which is very important when the goal is to prevent a single value ten years ahead with rather few data.

If we are to continue using this model, a few things will need to be considered:

- Some rather good fitting with a lower *lambda_1* value was found by accidentally ignoring one variable. Considering that, using Lasso instead of Ridge here or even applying some variable selection first before training our method would be relevant.
- An import analyse to proceed is to try to use the prior probability distribution in order to try to increase the importance of the closest year compared to the further as first considered.
- Obtaining a relevant uncertainty matching the reality is crucial for the final interpretation of our model.

## 5 Data visualisation

In order to easily visualise the forecast given by the different scenarios, we inserted a widget into a Jupyter notebook to render interactive plots. This allow us to change the explanatory variables values through the years we want to forecast consumption based on scenarios we set beforehand.

For each variable, we chose three possible linear evolution until 2030. A scenario for them is to stay constant from the last value of this variable (value of 2018).
To find the two other scenario we took the trend of the last 5 years of each variable and projected it to 2030. We approximated the percentage of the increased value from 2018 to the next year for the second scenario and inverted the percentage into a decrease for the third.
For the population however, instead of inverting the percentage, we doubled it as it is unlikely for a policy to reduce population.

### 5.1 Scenarios with linear regression.

Figure 13 shows a scenario where last known value of each constant is propagated until 2030.

Figure 14 represents a scenario where variables with low impact are set to increase. According to this model primary energy intensity makes almost no difference to the outcome. Gross domestic product is lowly and inversely related to the fossil energy consumption. Renewable energy has quite an impact but not sufficient if we follow the trend.

Figure 15 shows remaining variables (population, carbon intensity and nuclear energy consumption) have a real high impact in the final output of the model.

### 5.2 Scenarios with Bayesian regression.

The sole parameter having a noticeable impact on the output in this model is the population. (see fig. 16)

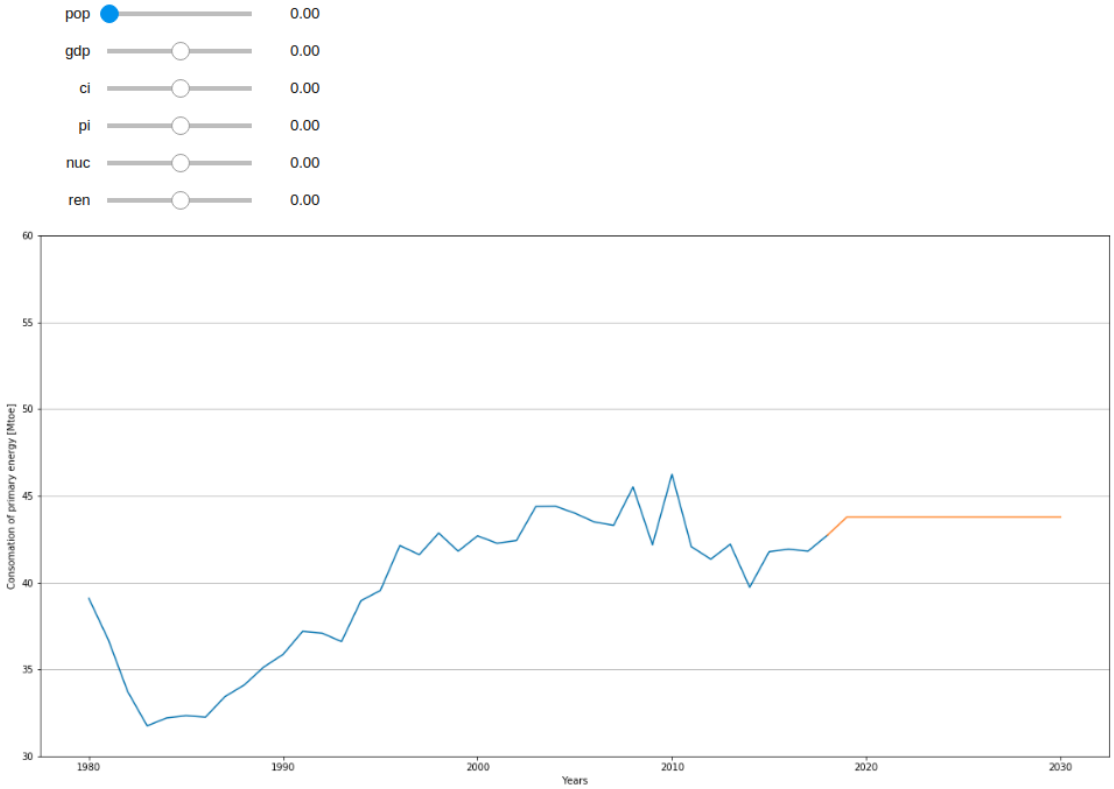


Figure 13

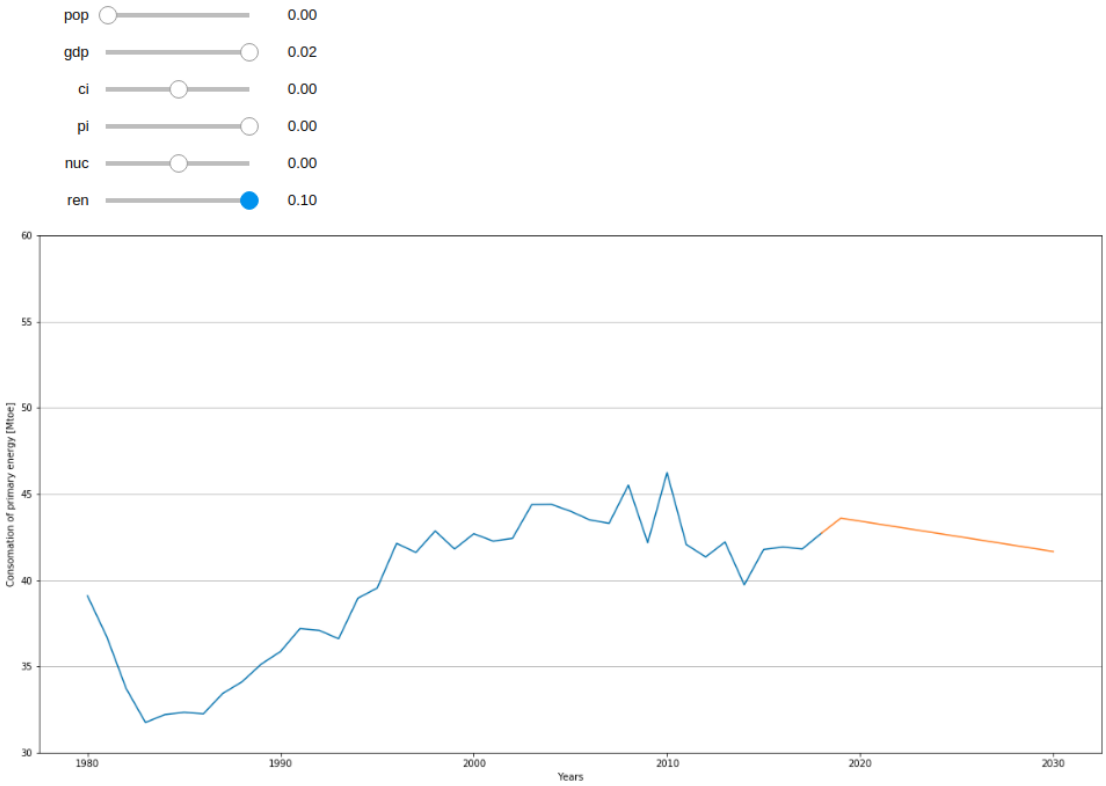


Figure 14

# 6 Annex

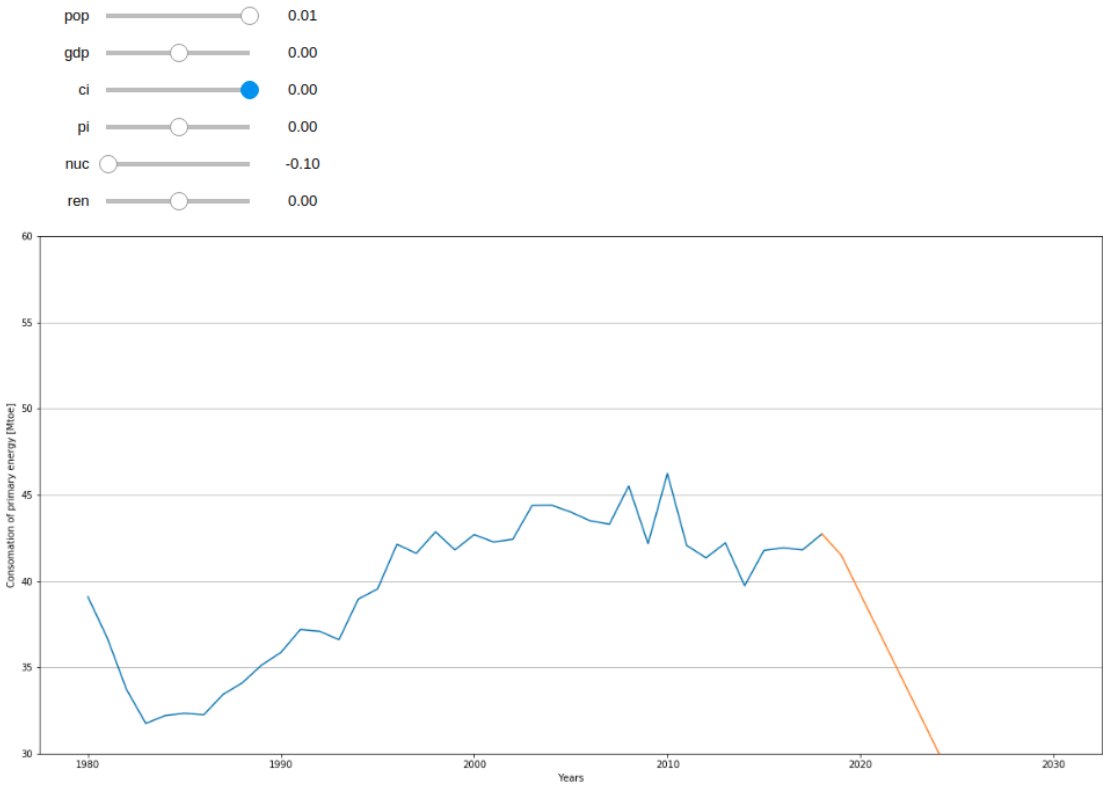


Figure 15

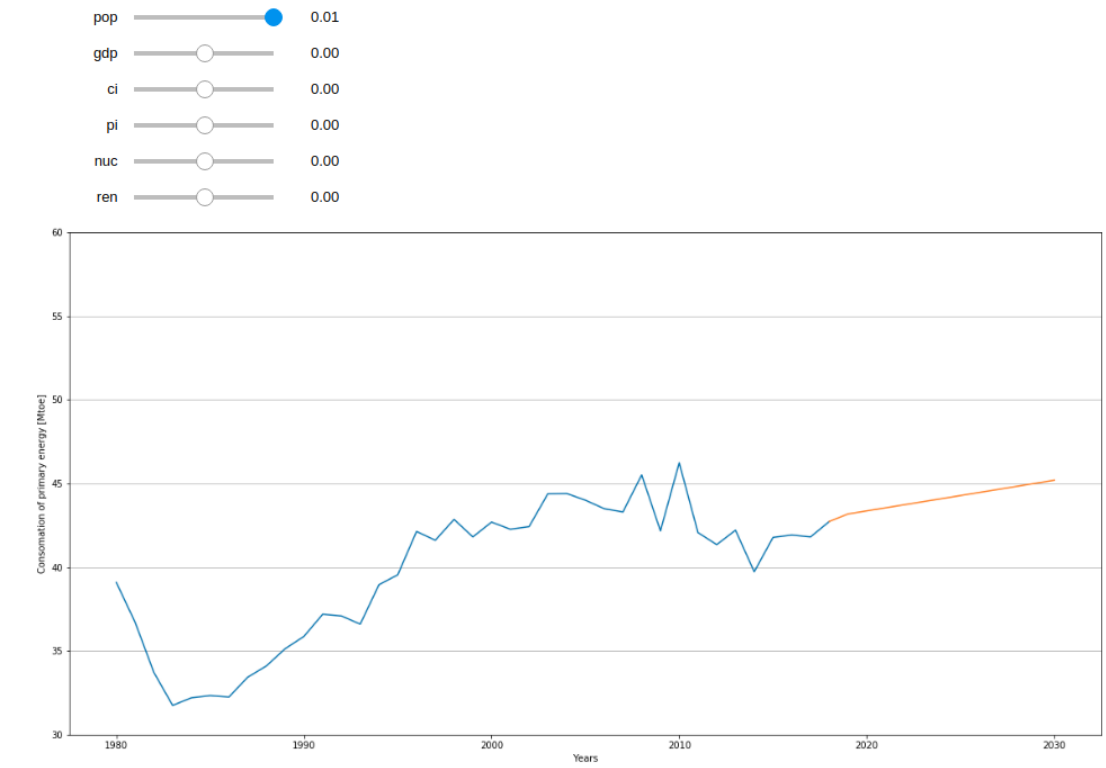


Figure 16

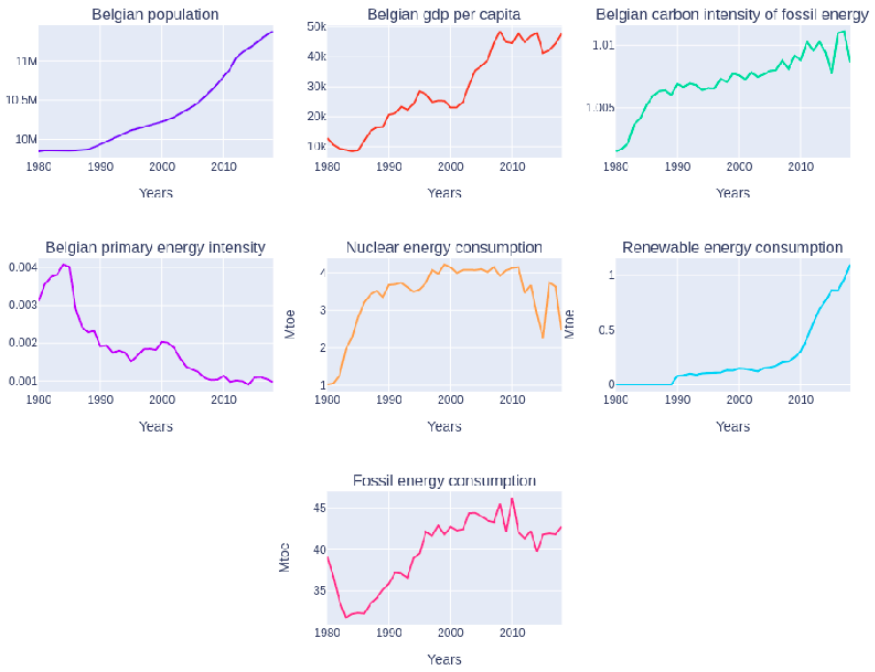


Figure 17: Scatter plot of independent variables in function of time