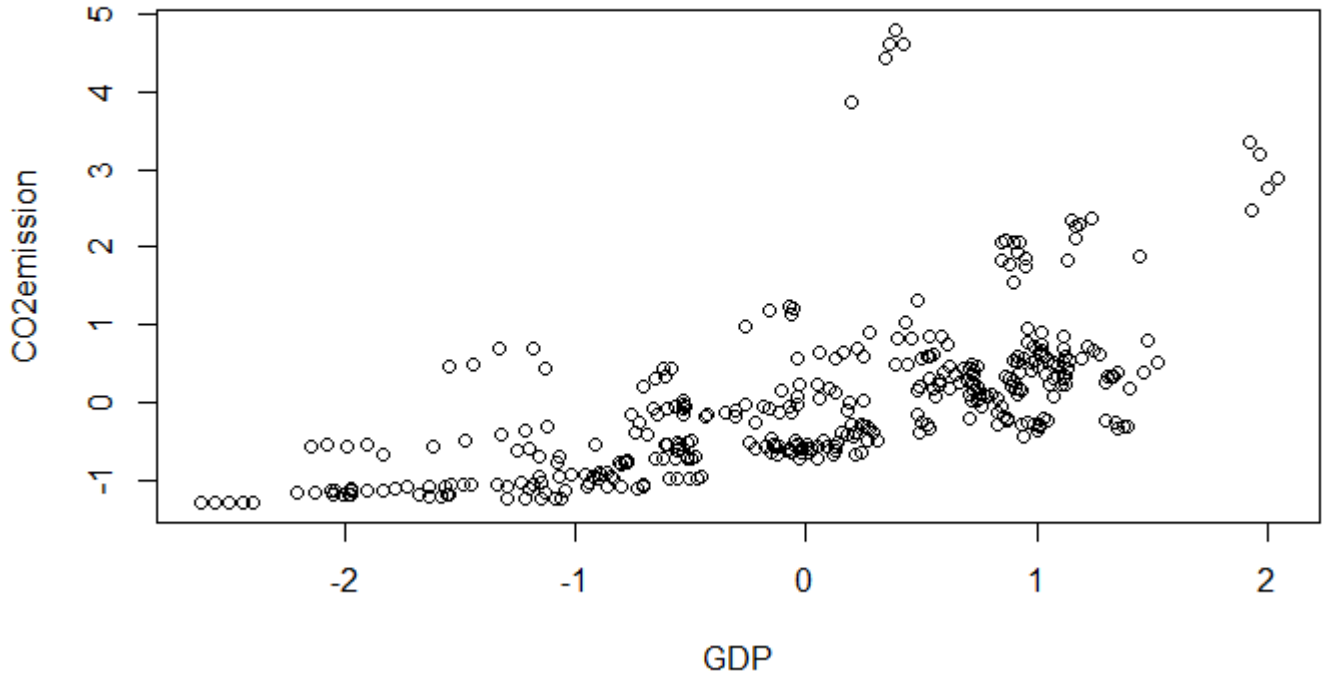
CO-2 dataset analysis

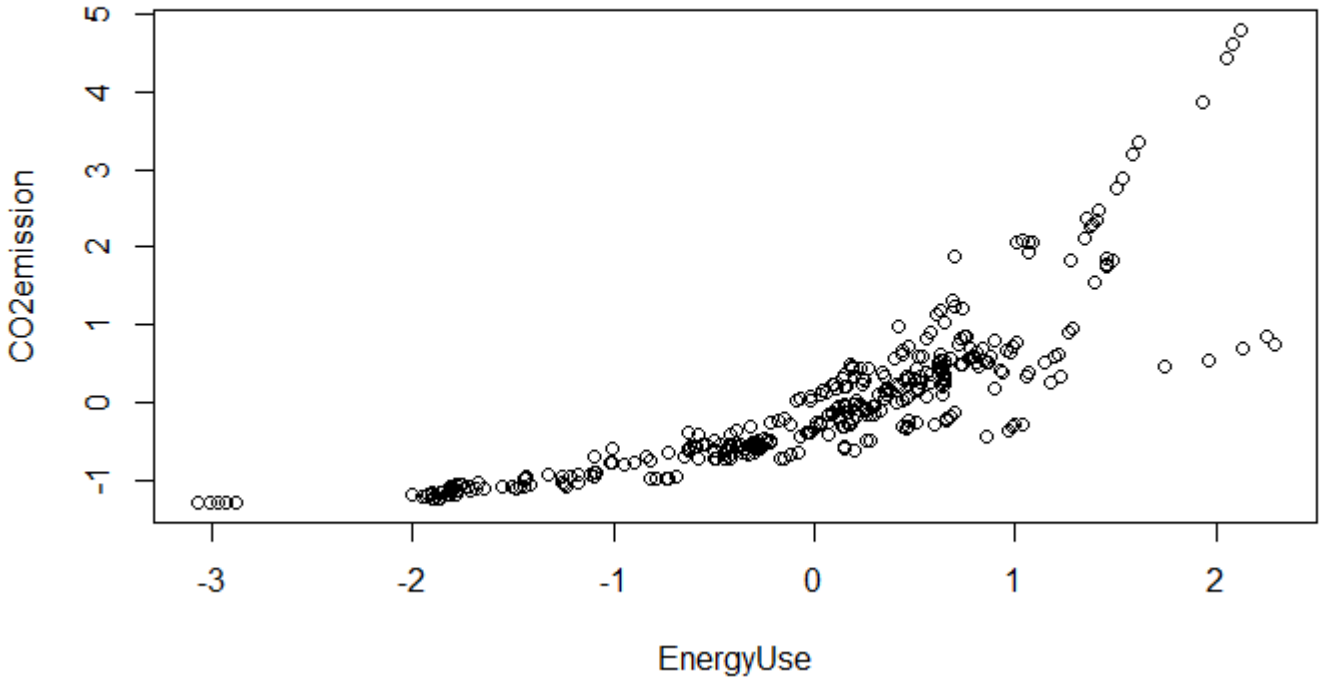
Bortolotti Simone, Franzè Lorenzo, Stancioi Ioana-Ruxandra

1. Task and dataset

Our goal is to observe if other factors in a country like: Country, Year (y), Energy use, GDP (Gross Domestic Product per capita), Population (pop), Low-carbon energy, Urban population (urb), Number of internet users (internet) can explain the CO2 emissions.

Our task is to explain the CO2 emissions with other variables, a second task consists in analyzing better the relation between GDP and CO2, in particular checking whether the richer we are, the more CO2 we emit, but at a certain point (threshold) this relation decays.





2. Preliminary analysis

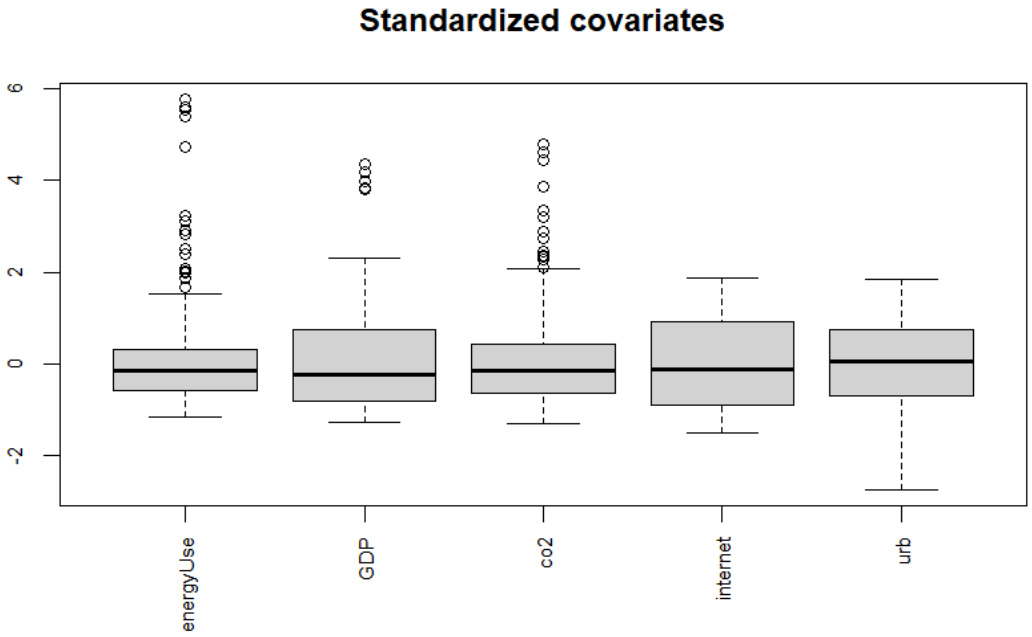
Our analysis focuses on per capita perspective, making it easier to compare countries with different population sizes,help to understand the average impact or behavior of an individual within a country and allow a fair comparison between countries regardless of their population sizes.

We applied some transformations on the data in order to consider the per capita perspective:

- Co2percap = unnormalized$co2percap,
- EnergyUse = normalized$EnergyUse * (1-unnormalized_data$Lowcarbon_energy/100),
- GDP = normalized$GDP,
- Pop = normalized$pop,
- Internet = unnormalized$internet/unnormalized$pop,
- Urb = unnormalized$urb/100

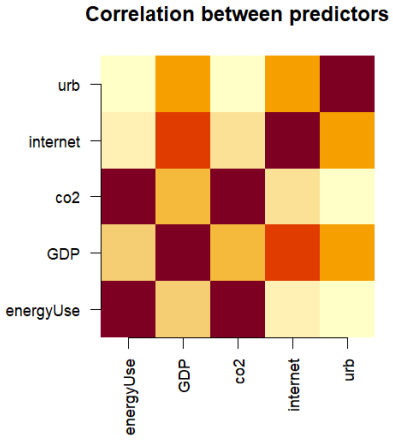
In particular Lowcarbon was used just to compute EnergyUse, but it was excluded from the following analysis. EnergyUse was considered as the amount of non-green energy produced by each country per capita, in the correlation analysis it can be seen as this new formulation of EnergyUse is way more correlated to the observation than when EnergyUse and Lowcarbon were both present.
While Urbanization was already a percentage, the Internet diffusion was not, therefore the ratio with the population was made.

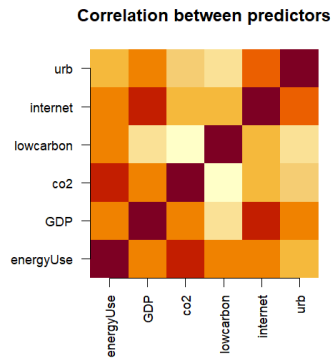
Given the high difference in their scale and distribution we normalized them, using ZScore and MinMax normalization following what was presumed better in each instance, in order to achieve maximal accuracy and interpretability.



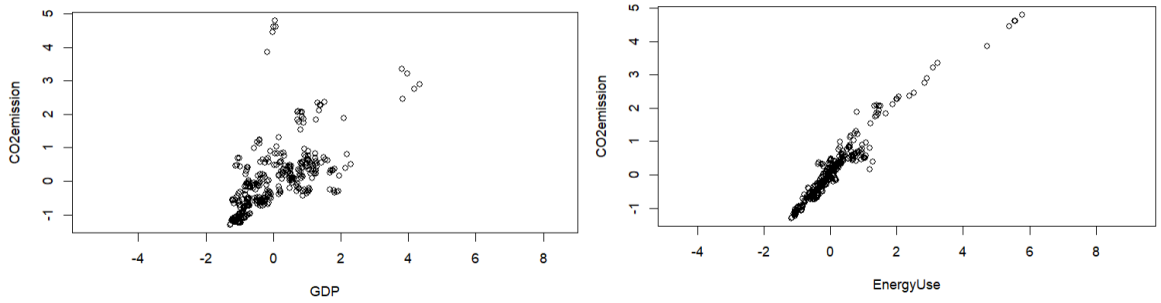
2.1. Correlation

On the matrix on the left we notice that our target distribution CO2 as expected has a strong positive correlation with energyUse and a middle correlation with GDP so they can be useful to predict the CO2 emissions. Interestingly, the internet variable is highly correlated with GDP. This correlation may slightly interfere with the linear regression model since we have a multicollinearity exactly with GDP, and this will make it difficult to determine the individual effect of each independent variable on the dependent variable.
The matrix on the right shows the relation with Lowcarbon, it has low correlation with other variables and for this reason besides the consideration above we have decided to include it in EnergyUse (by subtracting) and remove it from the datset.





Since the other variables are not highly correlated among them, we can proceed with the regression analysis. We plot the charts of co2 vs GDP and co2 vs energyUse and notice in Figure 6 that there are some outliers, we will not care where we use models robust to outliers, while we eliminate them in others.



3. BAS analysis and Model selection

Here we consider some Bayesian linear regression models using Zellner's g-priors and Jeffreys-Zellner-Siow (JZS) priors.

Using the classical approach we splitted the data into training and test sets. We studied how the Mean SSE changed, so we monitored both the bias and the variance of the models, paying attention not to overfit.
In particular, in JZS case we studied the importance of each coefficient and we found that the most important covariates are EnergyUse and GDP.

In the second part we added some covariates: energyUse^2 and GDP^2 finding that in this case the model that uses energyUse^2 and energyUse performs better both in terms of variance and bias.
We confirmed the same results also in JAGS (using LASSO).

4. Models (JAGS)

In order to better analyze the relationship between co2 and GDP we use JAGS to model more complex models that may help us to understand if at high incomes the relationship between co2 and GDP is still strong.

4.0 Inference models

We deployed different models with different assumptions.
In the first part we try to predict the co2 from other covariates, assuming in some cases to also know the distribution of the covariates, to see if we have improvements.
Then we try again using a Bayesian LASSO Prior to check for the importance of the covariates, and in this case only "urb" is discarded. This result is probability due to the middle correlation between internet and energyUse.

4.1 Informative models

In the first part of the analysis we tried to estimate the shape of the covariates in order to include them into the model and try to predict better the co2

4.2 Uninformative models

4.2.1 Threshold

In order to find out if the thesis that the CO2 per capita is proportional to the GDP, but only up to a certain threshold, we create models to explain the observations, both using this threshold and not, in order to confront them later.

Several trials have shown that putting a prior on the threshold creates oscillating results, therefore a simple normal model using just the GDP covariate and an intercept for both the group above and under the threshold has been created, with a uniform prior between 25000 and 35000 on the threshold and uninformative prior for the covariates, in order to fix the threshold.

	Mean	SD
Intercept_poor	1.9200	0.5547
Intercept_rich	11.0901	0.6738
GDP_poor	18.2226	3.2139
GDP_rich	-0.4352	0.9351
Std_dev	0.0586	0.0044
Threshold	27028.6787	843.6151

The threshold has been found around 27000. The procedure can be found in the code, using normalized data, but producing analogous results.

To further ensure our proposal for the threshold was fine even in front of the complete set of covariates, also a more complex model including all the covariates is included in the code, where the threshold posterior distribution has higher standard deviation, but is still around our value, making our result more robust. Another consideration is that there are not many data points around each value of GDP between 25000 and 35000, therefore a small variance in the result does not hinder the process.

4.2.2 Normal model with threshold

This model is built with uninformative flat priors, using the threshold of 27000 to separate in two groups the high-GDP and low-GDP countries. The observations are modeled using a simple normal model, where only the mean is influenced by the 4 covariates (even though urbanization and internet level did not belong to the best selected model, we tried several options and we included them in the model, to show how their results differ from the covariates that are mode useful).

$$
\begin{aligned}
\mu_i &\leftarrow \text{inprod}(X_i, \beta) \\
\text{co2percap}_i &\sim \mathcal{N}(\mu_i, \beta_\gamma) \\
\text{indicator}_i &\leftarrow (\text{unGDP}_i \geq \text{threshold}) \\
X_{i,1} &\leftarrow (1 - \text{indicator}_i) \\
X_{i,2} &\leftarrow \text{indicator}_i \\
X_{i,3} &\leftarrow \text{GDP}_i \cdot (1 - \text{indicator}_i) \\
X_{i,4} &\leftarrow \text{GDP}_i \cdot \text{indicator}_i \\
X_{i,5} &\leftarrow \text{EnergyUse}_i \cdot (1 - \text{indicator}_i) \\
X_{i,6} &\leftarrow \text{EnergyUse}_i \cdot \text{indicator}_i \\
X_{i,7} &\leftarrow \text{urb}_i \cdot (1 - \text{indicator}_i) \\
X_{i,8} &\leftarrow \text{urb}_i \cdot \text{indicator}_i \\
X_{i,9} &\leftarrow \text{internet}_i \cdot (1 - \text{indicator}_i) \\
X_{i,10} &\leftarrow \text{internet}_i \cdot \text{indicator}_i
\end{aligned}
$$

	Mean	SD
Intercept_poor	0.2955	0.32209
Intercept_rich	4.2510	0.70622
GDP_poor	1.1465	0.89081
GDP_rich	-0.3860	0.62585
EnergyUse_poor	47.2061	1.48681
EnergyUse_rich	45.6177	1.43180
Urbanization_poor	0.3562	0.59185
Urbanization_rich	-3.0046	0.93879
Internet_poor	-0.6484	0.59647
Internet_rich	-0.3522	0.70165

Quantiles	2.5%	25%	50%	75%	97.5%
Intercept_poor	-0.3479	0.08019	0.2968	0.51713	0.9343
Intercept_rich	2.8608	3.78712	4.2532	4.72457	5.6264
GDP_poor	-0.5971	0.54355	1.1480	1.76058	2.8487
GDP_rich	-1.6140	-0.80650	-0.3838	0.03251	0.8276
EnergyUse_poor	44.2673	46.23099	47.1938	48.20341	50.1317
EnergyUse_rich	42.8375	44.65093	45.5899	46.57545	48.4695
Urbanization_poor	-0.7912	-0.05304	0.3423	0.75384	1.5357
Urbanization_rich	-4.8460	-3.63380	-3.0098	-2.36049	-1.1942
Internet_poor	-1.7998	-1.04570	-0.6597	-0.25376	0.5356
Internet_rich	-1.7177	-0.81962	-0.3394	0.10201	1.0681

The results partially confirm our initial thesis: the parameter associated to the GDP of low-GDP countries shows a direct proportionality to the CO2 per capita, while the relation is near zero for the high-GDP group, even though the zero value is included also in the 95% credible region of the parameter associated to the GDP of low-GDP countries (in the picture the 80% credible region is used, to keep outside the zero axis and show that it is still improbable that the covariate is irrelevant). It is also visible how EnergyUse is by far the covariate

that explains the largest part of the variability in the dataset, indeed it is more than an order of magnitude bigger than the others both over and under the threshold (this can be a sign of relevance since all covariate have been normalized).

Regarding the Urbanization and Internet covariate, they are as expected near zero, with the exception of Urbanization over the GDP threshold: this should inversely relate the Urbanization of a high-GDP country with its pollution, this could mean that living in a city could be more efficient from the pollution point of view. Finally the intercept parameter is higher for the countries over the GDP threshold, this was expected by our thesis since it takes on itself the CO2 per capita portion that was explained by the GDP covariate in the data points under the threshold.

4.2.3 Normal model without threshold

To consider the same model without a threshold it is enough to move the threshold to 1000, in order to not create two groups, but just one. The results are:

	Mean	SD
Intercept	0.7381	0.2775
GDP	0.9992	0.4996
EnergyUse	47.3396	1.0690
Urbanization	-0.4621	0.5205
Internet	-0.0404	0.4557

Clearly without using the threshold we cannot hope to find a difference between Low-GDP and High-GDP countries, the proportionality is maintained but with a lower parameter.

4.2.4 Gamma model with threshold

A second general model, with equally uninformative prior and a fixed threshold, but using a gamma model, was created, to check the findings of the first one and its robustness. The results, reported below, brought to similar consideration with respect to the ones for the original model, therefore only the original model was considered later.

	Alfa	Beta
Intercept_poor	0.3393	0.4297
Intercept_rich	8.1026	2.2230
GDP_poor	1.2990	0.9797
GDP_rich	-0.0034	0.9600
EnergyUse_poor	80.6386	5.3695
EnergyUse_rich	66.2983	5.7611
Urbanization_poor	2.4925	1.0492
Urbanization_rich	-1.0524	3.0609
Internet_poor	0.3217	1.2804
Internet_rich	-0.2116	2.2215

Quantiles	2.5%	25%	50%	75%	97.5%
Intercept_poor	-0.4463	0.03149	0.32752	0.6452	1.170
Intercept_rich	3.5053	6.62107	8.16108	9.6720	12.111
GDP_poor	-0.5678	0.63269	1.27294	1.9569	3.269
GDP_rich	-1.8720	-0.65513	-0.02142	0.6361	1.892
EnergyUse_poor	70.0574	76.97383	80.59682	84.4011	90.849
EnergyUse_rich	55.5711	62.28592	66.13402	70.2348	77.795
Urbanization_poor	0.4706	1.78305	2.46256	3.2063	4.573
Urbanization_rich	-6.9383	-3.04771	-1.10059	0.8323	5.600
Internet_poor	-2.1877	-0.56244	0.30464	1.1633	2.843
Internet_rich	-4.6210	-1.56529	-0.26318	1.1909	4.315

4.3 Time-series model

Considering the past observation as covariates we can greatly improve the prediction in our situation, since the CO2 per capita of a whole country is a value not prone to peaks and sudden changes. Since each country has 5 points associated (with the exception of Belarus, Hong Kong, United Arab Emirates, which are excluded from the dataset for this section), corresponding to the period 2005-2009 we will be able to create a model which only focus on the last observation and the main covariates, to obtain 4 periods for each country. It must also be noticed that the first two predictions (2006-2007) refer to a standard period, while the other two are taken during the 2008 world economic crisis. This means the prediction will likely have less accuracy from a period to the other, and more variability could be present in the later years.

$$Y[i] \sim N(\mu[i], \tau)$$
$$\mu[i] \leftarrow \alpha \cdot Y[i-1] + m_1 \cdot x_1[i-1] + m_2 \cdot x_2[i-1] + m_6 \cdot (x_1[i] - x_1[i-1]) + m_7 \cdot (x_2[i] - x_2[i-1])$$

$$\alpha \sim \text{Uniform}(-1.5, 1.5)$$
$$\tau \sim \text{Gamma}(0.1, 10)$$
$$m_1 \sim N(0.0, 1.0 \times 10^{-4})$$
$$m_2 \sim N(0.0, 1.0 \times 10^{-4})$$
$$m_6 \sim N(0.0, 1.0 \times 10^{-4})$$
$$m_7 \sim N(0.0, 1.0 \times 10^{-4})$$

The model shows how the last observation contributes with more than 90% of its original value to the prediction, with the covariate representing the difference between the EnergyUse of the current and the past year, called m6, which is the only one that has not the zero axis in its 95% credible interval.

	mean	sd
alpha	0.9414781	0.04869728
sigma2	0.3395359	0.04511665
m1	2.2496952	2.96750236
m2	0.2059677	0.61487499
m6	41.1882197	5.90675886
m7	4.8965022	5.38595143
deviance	155.5459971	9.11750106

The next step done is to separate the dataset in two parts, assuming that more variability could be present when the economic crisis happened. The first model uses period 2005-2006 to predict 2006-2007, while the second uses 2007-2008 to predict 2008-2009.

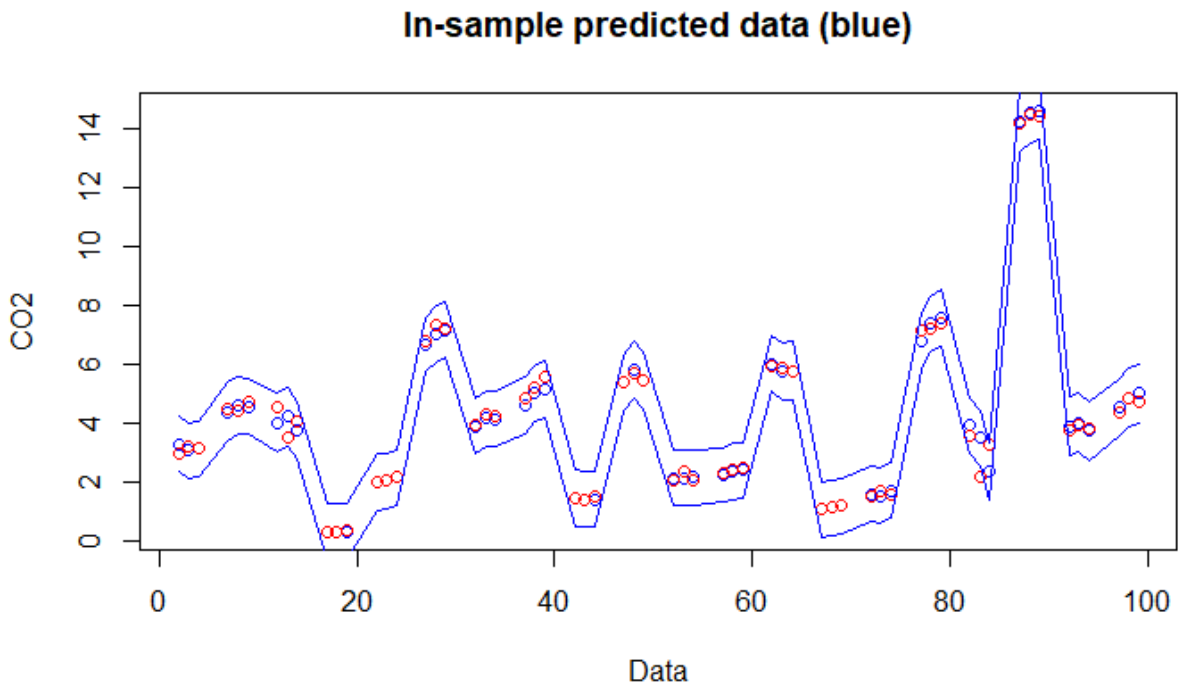
The two new models confirm in a way what was expected:

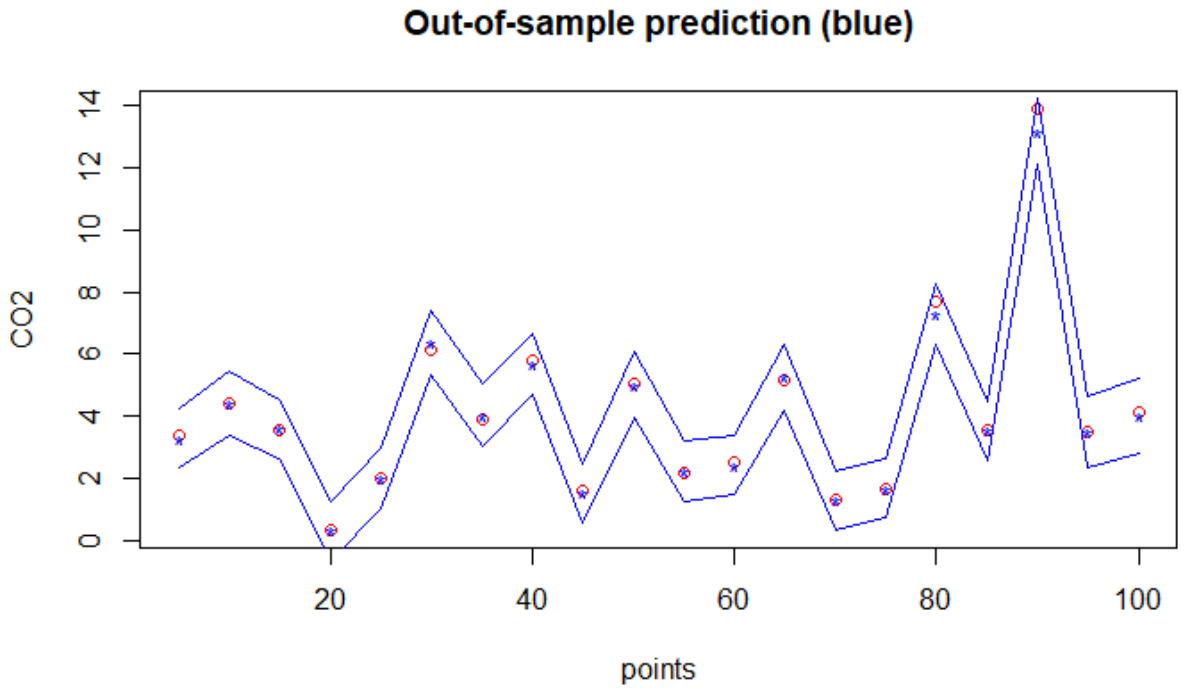
- Alpha has a lower standard deviation in the first period, since it sees data from a period where the situation was more stable and the prediction of the year before was more useful.
- Sigma2 is lower in the first period, since the mean of the predicted observation is predicted in a more accurate way when the crisis is not present.
- Both m1, m6 (which are the parameters related to the energy use) have an higher variance in the "crisis period" and the importance of the difference in energy use decreases a lot.
- Instead m7 increases in relevance during the crisis, but it has a variance which is extremely high, with a credible region which comprehends 0. This probably derives by the fact that in the crisis period the GDP-difference is somewhat used to understand how much the crisis impacted.
- Lastly also the Deviance, inversely correlated to the likelihood of the model of explaining the data, is lower in the standard period.

05-07	mean	sd	07-09	mean	sd
alpha	0.9305	0.0926	alpha	0.9355	0.1138
sigma2	0.6387	0.1582	sigma2	0.7781	0.1909
m1	3.1174	5.9042	m1	1.2686	7.0215
m2	0.0215	1.5218	m2	0.8415	1.6412
m6	44.1417	12.5474	m6	14.1442	15.6074
m7	4.9534	20.3149	m7	20.1102	17.4323
deviance	64.0355	9.3684	deviance	77.8552	8.0963

Even if the data contribute to create 2 somewhat different models with respect to the periods 2005-2007 and 2007-2009, the fact that the time series are long only 5 data points and the fact that the co2percapita of a nation is usually a stable value, means that the prediction results to be satisfactory even with few points, also using just the period 2005-2006 to predict the period 2006-2009.

The following are the In-sample and Out-of-sample prediction of the first model (period 2005-2008 used to predict year 2008-2009), this last with the computation of the Mean Square Error and the R squared index, all the other are not included since all graphs are pretty similar, as it can be seen by the table of the MSE and R squared below:





Here the MSE and R2 there are reported below the period used for training and the predicted period, it can be noticed all combinations of MSE and R2 describe similarly satisfactory levels of uncertainty:

	05-06 06-09	05-07 07-09	05-08 08-09
MSE:	0.439	0.551	0.924
R2:	0.986	0.982	0.969

	05-06 06-07
MSE:	0.173
R2:	0.994

	05-06 07-08	05-07 07-08
MSE:	0.154	0.150
R2:	0.995	0.995

	05-06 08-09	05-07 08-09	05-08 08-09
MSE:	0.990	0.952	0.924
R2:	0.967	0.968	0.969

Finally, the last point is introducing a threshold. Trying to separate the richest and poorest countries, the results are indeed more accurate than with all data together, it was also visible that the difficulties in predicting the crisis period data are still present. This difficulties could be increased by the small number of data points, unable to explain the variability in a context where a switchpoint, several covariates and a threshold are present.

Both accuracy found trying to predict only a portion of the countries is actually higher than the accuracy we had using all datapoints, even if this happens without any parameter changing its sign or its magnitude. This could mean that the two groups present different characteristics and in particular since the High-GDP groups present the highest accuracy, we could think there is less variability in the general CO2 per capita value once the threshold is reached (even if the GDP and GDP-difference parameter by themselves are not far from zero).

```
All-data 05-08 to predict 08-09:
MSE:  0.924
R2:   0.969

High-GDP 05-08 to predict 08-09:
MSE:  0.041
R2:   0.995

Low-GDP 05-08 to predict 08-09:
MSE:  0.070
R2:   0.995
```

```
Low-GDP 05-08 to predict 08-09:
                    mean         sd
alpha         0.97281260 0.01882084
sigma2        0.46537691 0.07556155
m1            1.25447141 1.45214557
m2            0.07163004 0.51977068
m6           30.77443218 6.49935315
m7            4.10598720 4.49939565
```

```
High-GDP 05-08 to predict 08-09:
                    mean         sd
alpha         0.93706022  0.03886073
sigma2        0.23362421  0.03112142
m1            2.35499572  1.37874488
m2            0.03115735  0.19312618
m6           32.27048530  4.24668766
m7            0.80912565  2.12125569
```

5. Clustering

Since it may be useful to look into clustering, in order to find if some regularities are hidden in our data, three levels of analysis have been carried out: at first we tried to extract groups from single covariates, then some 2-dimensional and 3-dimensional models were tried, but this last without any significant result. Indeed the 3-dimensional model was only able to extract the outliers, while if those were excluded they simply created one big group containing all data points.

The simpler, 1-dimensional, models were somewhat useful in finding some deeper information in the data, but their findings are somewhat already visible in the histograms, therefore only the CO2 per capita clusters will be described.

$$
\begin{aligned}
&\# \text{ Likelihood} \\
&z_i \sim \text{Categorical}(w) \\
&y_i \sim N(\mu_{z_i}, \sigma^2_{z_i}) \\
\\
&\# \text{ Prior} \\
&\mu_i \sim N(0, 0.01) \\
&\sigma^2_i \sim \text{Gamma}(0.01, 0.01) \\
\\
&w \sim \text{Dirichlet}(a)
\end{aligned}
$$

There are 3 clusters, of different value and number of points: the first group contained the most variability, it accounted for almost one seventh of the data points and contained the group of lowest polluting countries, with a mean of 1.25. The middle-polluting countries group was instead way more concentrated at a mean level of 7.24 and it contained more than four countries out of fifth. The remaining group was the heaviest-polluting, containing only around 5% of the countries, with a mean of 18.87, almost three times the middle group. The large difference in the values observed between the top polluters (like Canada, Australia and the USA) and the average group shows that a small percentage of the world population is responsible for a more relevant percentage of the pollution.

CO2percapita	Mean	SD
mu[1]	1.2560	0.10527
mu[2]	7.2472	0.21785
mu[3]	18.8770	0.51762
sigma[1]	3.7946	1.18170
sigma[2]	0.1009	0.01014
sigma[3]	0.6410	0.28479
w[1]	0.1345	0.02395
w[2]	0.8170	0.02662
w[3]	0.0485	0.01255

EnergyUse	Mean	SD
mu[1]	4935.5858	127.6973
mu[2]	23255.0733	1371.3197
mu[3]	44316.3451	5078.1127
sigma[1]	0.0000	0.0000
sigma[2]	0.0000	0.0000
sigma[3]	0.0000	0.0000
w[1]	0.1052	0.0182
w[2]	0.5994	0.0935
w[3]	0.2953	0.0929

GDP	Mean	SD
mu[1]	8681.4802	1015.8235
mu[2]	20042.7215	1823.9596
mu[3]	42230.8380	1819.0478
sigma[1]	0.0000	0.0000
sigma[2]	0.0000	0.0000
sigma[3]	0.0000	0.0000
w[1]	0.3368	0.0695
w[2]	0.1855	0.0727
w[3]	0.4776	0.0434

The 2-dimensional model correlating CO2 per capita and GDP is more meaningful, in particular the two groups found have average GDP around 23000 and 40000, with the first group accounting for more than three times the point in the second group. The second group is also responsible for an average CO2 per capita production of 12.5, way higher than the 5.5 of the first group.

$$z_i \sim \text{Bernoulli}(w)$$
$$y_{i,1:2} \sim \text{Multivariate Normal}\left(\mu_{z_i+1,1:2}, \sigma_{z_i+1,1:2,1:2}\right)$$

$$\mu_{1,1} \sim N(0, 0.1)$$
$$\mu_{1,2} \sim N(0, 0.1)$$
$$\mu_{2,1} \sim N(0, 0.1)$$
$$\mu_{2,2} \sim N(0, 0.1)$$
$$w \sim \text{Beta}(1, 1)$$

$$\sigma^2_{1,1,1} \sim \text{Gamma}(1, 1)$$
$$\sigma_{1,1,2} \sim N(0, 0.1)$$
$$\sigma_{1,2,1} = \sigma_{1,1,2}$$
$$\sigma^2_{1,2,2} \sim \text{Gamma}(1, 1)$$

$$\sigma^2_{2,1,1} \sim \text{Gamma}(1, 1)$$
$$\sigma_{2,1,2} \sim N(0, 0.1)$$
$$\sigma_{2,2,1} = \sigma_{2,1,2}$$
$$\sigma^2_{2,2,2} \sim \text{Gamma}(1, 1)$$

	Mean	SD
mu[1,1]	5.5809	0.3823
mu[2,1]	12.2621	1.5298
mu[1,2]	0.2933	0.0275
mu[2,2]	0.4997	0.0528
w	0.2220	0.0723

6. Conclusions

Analyzing the data we were given we found several interesting relationships between the observations and some covariates. The main point of interest was the relationship between CO2 emissions and GDP, which we were in a way able to show as non linear, presenting a threshold of GDP per capita, around 27000$. The threshold position didn't change too much according to the assumptions we make about the influence between CO2 and GDP and whether we consider the intercept or not, therefore we considered it fixed in our analysis

While below this threshold the relation was found as proportional, any relation searched over the threshold was weak, as if the GDP stopped influencing the CO2 production.

Having tried several ways to build a model, we found the strongest to be by far the one using autoregression, which in a timespan so short as the one present in our data (5 years) was able to predict in a satisfactory way the future data points, showing how the previous year observation was in a way the most important covariate. We noticed also that the results were easier to predict for High-GDP countries, this could be seen as a sign that these countries have a less variable CO2 production. Another distinction we made was between the points before and after the 2008 economic crisis, after which the data are more complex to predict and less tied to the previous data.

Between the remaining covariates, the one with a strongest relation to the CO2 production was EnergyUse, which we defined as the amount of non green energy produced by a country, which has proved to be more useful of both the separate data in the initial dataset, measuring the total energy produced and the total green energy produced. Less useful were instead the Internet diffusion, Urbanization and Population, even if this value was useful in computing the per capita values.