Bayesian Statistical Analysis of CO2 Emissions Data

Bayesian Learning and Montecarlo Simulation

Gabriele Carrino
Gabriele Carminati

POLITECNICO
MILANO 1863

Contents

1 Introduction

Human emissions of carbon dioxide and other greenhouse gases are a primary driver of climate change and present one of the world's most pressing challenges. To understand this complex issue better our analysis will focus on:

- **Long-Term Trends**: we will examine how CO2 emissions have evolved over the past century, identifying key periods of increase or decrease in the speed of emissions and correlating these with historical events.
- **Recent Factors Influencing CO2 Emissions**: we will analyze the impact of various factors on CO2 emissions from 2006 to 2009. This includes investigating the relationship between CO2 emissions and variables such as energy consumption, GDP and the adoption of low-carbon energy sources.

2 Datasets

Brief description of the datasets:

- **Dataset 1** (Long Term) encompasses annual CO2 emissions for each year from 1900 to 2022. This dataset provides a comprehensive perspective on CO2 emissions over more than a century, allowing us to identify long-term trends and patterns.
- **Dataset 2** (Recent Factors) is more granular and concentrates on the years 2006 to 2009. It includes various features such as energy use per capita, GDP, population, CO2 emissions per capita, the percentage of low-carbon energy in total energy production, urbanization levels, and internet usage. This dataset allows for a detailed analysis of specific factors that may influence CO2 emissions during a recent, focused timeframe.

By integrating these two datasets, we aim to analyse CO2 emissions over time and identify the factors driving changes in recent years.

3 Long-Term Trends

We begin our analysis by considering the long-term trends in annual CO2 emissions over the last century. By plotting the Annual CO2 emissions on a real plane we can spot an **exponential trend**. For this reason it could be interesting to analyse the **logarithm** of the data and its **increment** to see if we can find some interesting results.

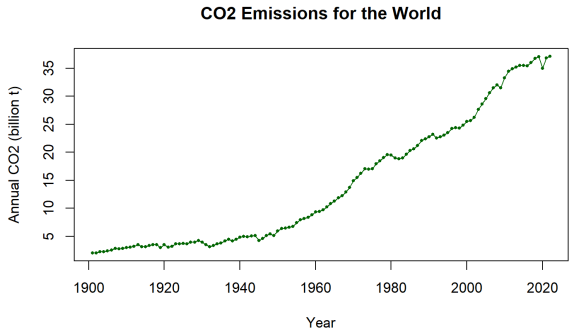

Figure 1: Annual CO2 emissions in the world

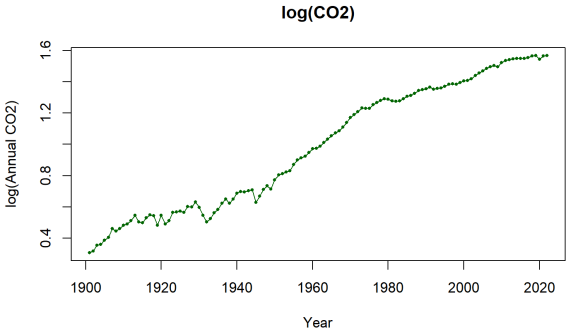

Figure 2: log(CO2) emissions

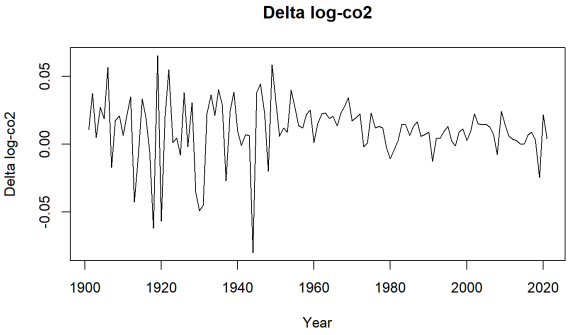

Figure 3: Delta log(CO2)

3.1 Change point model for the variance

From the delta log-CO2 plot it is clear that some kind of **shift in the variance** is happening. To analyse this in more detail we will fit a change point model on the variance to our data. We assume for our data a Gaussian likelihood:

$$y_t \sim \mathcal{N}(\mu_0, \sigma_t^2).$$

where:

$$y_t := \log(x_{t+1}) - \log(x_t)$$

An event occurs at a random unobserved time τ, the so-called "change point". This is reflected in the assumption:

$$\sigma_t^2 = \sigma_1^2 \quad t < \tau$$

$$\sigma_t^2 = \sigma_2^2 \quad t \geq \tau$$

where σ_1^2, σ_2^2 are unknown.

The model is as follows:

$$
\begin{aligned}
y_t &\sim \mathcal{N}(\mu_0, \sigma_t^2) \\
\sigma_t^2 &= \sigma_1^2\{t < \tau\} + \sigma_2^2\{t \geq \tau\} \\
\tau &\sim \mathcal{U}(0, M) \\
\mu_0 &\sim \mathcal{N}(0, 100) \\
\frac{1}{\sigma_i^2} &\sim \mathcal{G}(0.001, 0.001), \quad i \in \{1, 2\}
\end{aligned}
$$

M is a parameter that indicates the number of datapoints. We choose the priors for μ_0 and $\frac{1}{\sigma_i^2}$ to be weakly informative.

3.1.1 Plots & Results

Our analysis suggests a potential shift in the variance of the emission around **1951**. With 90% confidence, we can say this shift likely occurred between 1949 and 1956. This timeframe is particularly interesting because it coincides with several major events that could be linked to this change in trend:

- **The Marshall Plan (1948)**: This large-scale U.S. program aimed to rebuild Europe after World War II. It likely spurred economic activity and potentially increased CO2 emissions.
- **The Post-War Economic Boom (1945-1973)**: This period of rapid economic growth across the globe could have significantly contributed to rising emissions.

One possible interpretation of these findings is that improved economic conditions led to both a decrease in the variability and an increase in the overall quantity of emissions. This suggests a shift towards more consistent, but higher, emissions during this time.

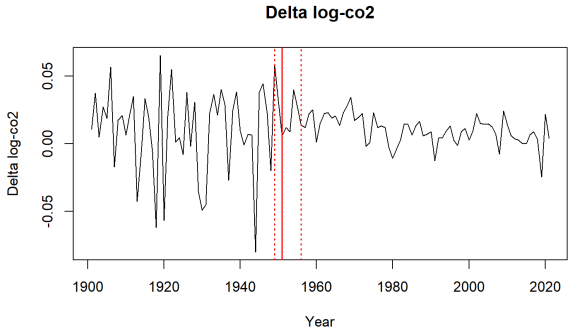

Figure 4: Change point in the variance

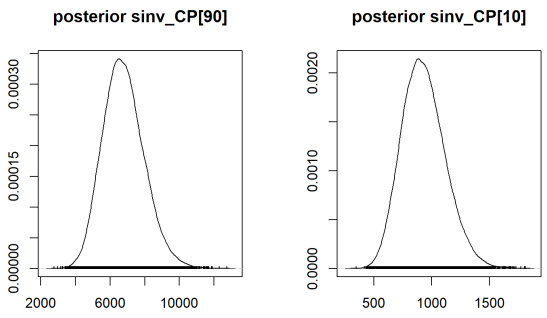

Figure 5: Posterior distributions of $\frac{1}{\sigma_i^2}$

3.2 Change point model for the mean

Given the potential shift in CO2 emission patterns around 1951, a next step could be to investigate if there are further changes in trend following this period. The Post-War Economic

Boom ended with a recession (1973-1975), followed by an even deeper one in the early 1980s. This historical context prompts the question: **Can we detect another change point in the CO2 emissions data after 1951?**

For the next analysis we will try to fit the data after the first shift (1951) with a change point model on the mean of the delta log-co2 in the years after 1956 (95% chance the shift in variance has already happened, credible interval). This means that we will analyse the change in speed of the increase of CO2 (mean of the logarithm) and not its variance anymore.
The model is as follows:

$$
\begin{aligned}
y_t &\sim \mathcal{N}(\mu_t, \sigma_0^2) \\
\mu_t &= \mu_1\{t < \tau\} + \mu_2\{t \geq \tau\} \\
\tau &\sim \mathcal{U}(0, M) \\
\mu_1 &\sim \mathcal{N}(5, 100) \\
\mu_2 &\sim \mathcal{N}(0, 100) \\
\frac{1}{\sigma^2} &\sim \mathcal{U}(0, 100000)
\end{aligned}
$$

M is a parameter that indicates the number of datapoints. We choose the priors for μ_0 and $\frac{1}{\sigma^2}$ to be weakly informative.

3.2.1 Plots & Results

As we can see from the plots we can indeed spot a change in the mean of our model around **1973**, with a **90%** chance this change occurred **between 1970 and 1978**.

An interpretation of this finding could be that the worsened economic conditions caused by the economic recessions of that era may have contributed to a slowdown in CO2 emissions growth.

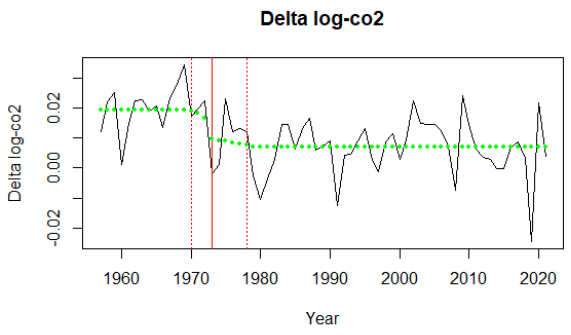

Figure 6: Change point in the mean

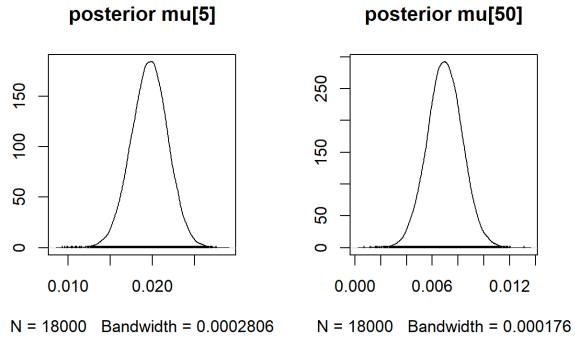

Figure 7: Posterior distributions of μ_i

3.2.2 Splitting the data in three periods

Our analysis identified two potential turning points in the CO2 emission trend: 1951 and 1973. To gain a clearer visual understanding of these shifts, we'll split the data into three periods:

before, between, and after these critical years. By fitting regression lines to the log of CO2 emissions for each period, we can directly visualize the changes in the emissions trajectory.

As we can see from Figure 8 and 9 there is a clear shift in the speed of increase of emissions before and after the change points.

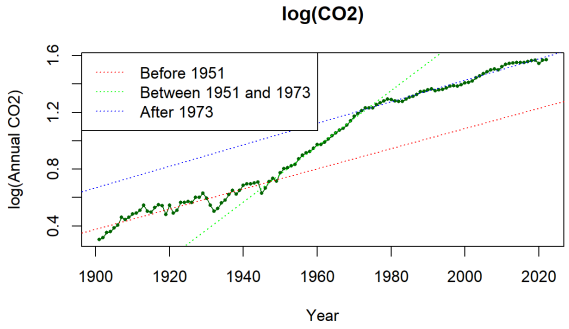

Figure 8: regressions on log(CO2)

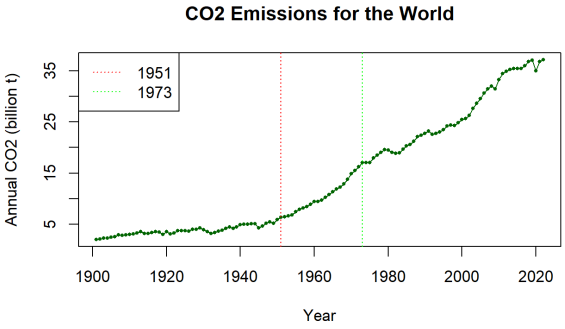

Figure 9: Change points on the real plane

3.3 AR(1) model - Time series prediction

Building on the identified change points in CO2 emissions, a crucial next step is to explore what lies ahead. The current estimate (2023) for the world's remaining carbon budget for a 50% chance to stay under 1.5 °C (2.7 °F) is 250 gigatonnes CO2. **Can we predict when we might reach a critical level of CO2 emissions that would push global temperatures 1.5°C above pre-industrial levels?**

We will carry on this analysis by fitting an **AR(1)** model with parameters μ, α and σ^2 to the delta of the logarithm of CO2 emissions on data after 1978 (95% that the change point has happened).

We assume the following priors:

$$\alpha \sim \mathcal{U}(-1, 1)$$
$$\mu \sim \mathcal{N}(0, 10)$$
$$\tau = \frac{1}{\sigma^2} \sim \mathcal{G}(0.001, 0.001)$$

We choose the priors for μ and $\frac{1}{\sigma^2}$ to be weakly informative.

3.3.1 Plots & Results

Based on our analysis, continuing the current emissions trajectory could push us towards 1.5°C warming by 2029.

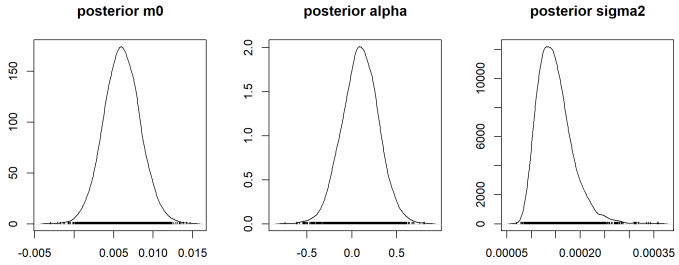

Figure 10: AR(1) posteriors

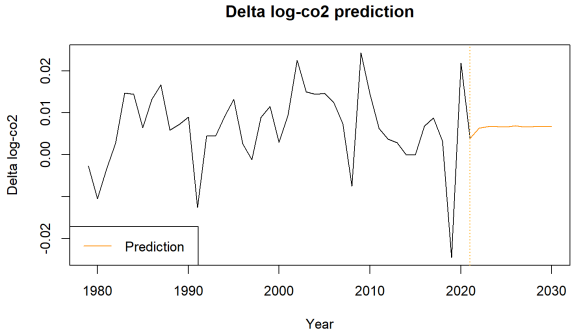

Figure 11: Delta log prediction

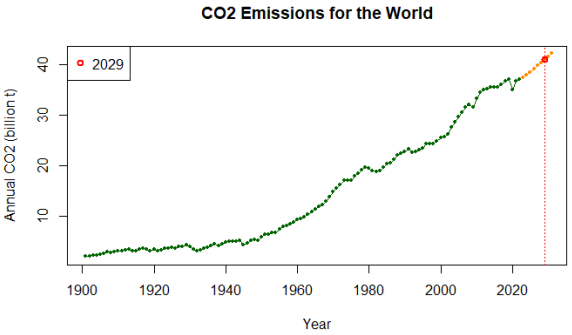

Figure 12: Prediction on real plane

3.4 White Noise Model with switching variance

In our previous analyses, we have examined the trends in the increase of CO2 emissions. Now we want to ask a different question: **can we detect global crisis and major worldwide events by looking at CO2 emissions alone?**

Do they cause a noticeable variance in the trend of CO2 emissions?

We will carry on this analysis by fitting a white noise model with switching variance on the increment of the log of CO2 emissions.

We will split our data by leveraging the change points discovered in two sub-datasets: years **before 1949** (5% quantile of the change point on the variance) and **after 1978** (95% quantile of the change point on the mean).
The model is as follows:

$$x_t := \log(y_{t+1}) - \log(y_t)$$

and we assume that

$$x_t \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_t^2)$$

with

$$\sigma_t^{-2} = \beta_1 + \beta_2 \delta_t \qquad \delta_t \sim Ber(p)$$

We assume the following priors:

$$\begin{aligned}
\beta_1 &\sim \mathcal{G}(0.001, 0.001) \\
\beta_2 &\sim \mathcal{G}(0.001, 0.001) \\
\delta_t &\sim Ber(p) \\
p &\sim Beta(1, 1) \\
\mu &\sim \mathcal{N}(0, 10)
\end{aligned}$$

We choose the priors for μ, β_1, β_2 and p to be weakly informative.

3.4.1 After 1978

We start our analysis with the years after 1978

3.4.2 Preprocessing

To make the results more accentuated we applied some transformations to our data. In particular we normalized and then elevated to power 3. This makes outliers in the data easier to spot.

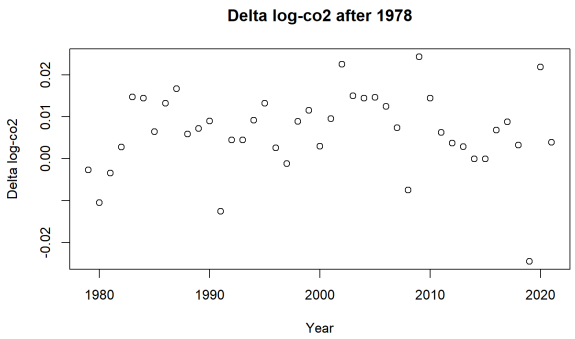

Figure 13: Data before preprocessing

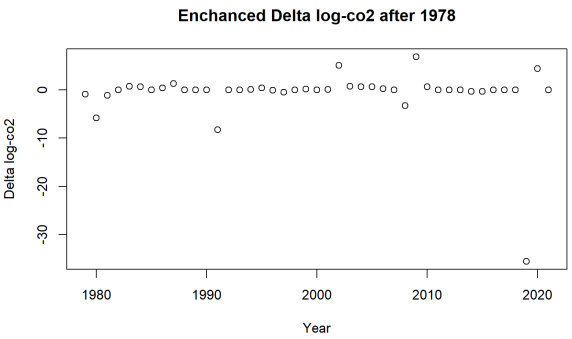

Figure 14: Data after preprocessing

3.4.3 Plots & Results

The years with the most significant variance in CO2 emission trends are: **1980**, **1991**, **2002**, **2008**, **2009**, **2019**, **2020**. Those years indicate periods where the subsequent year experienced a noticeable change in CO2 emissions (we analysed the delta of the data) and can be directly linked with major global economic phenomena:

- 1980: **Early 1980s Recession**, decrease in CO2 emissions.
- 1991: **Dissolution of the Soviet Union** and **Early 1990s Recession**, decrease in CO2 emissions.
- 2002: **Introduction of the Euro**, **11/09/2021 attacks**, increase in CO2 emissions.
- 2008/2009: **Global Financial Crisis**, decrease in CO2 emissions.
- 2019/2020: **Covid-19 Pandemic**, decrease in CO2 emissions.

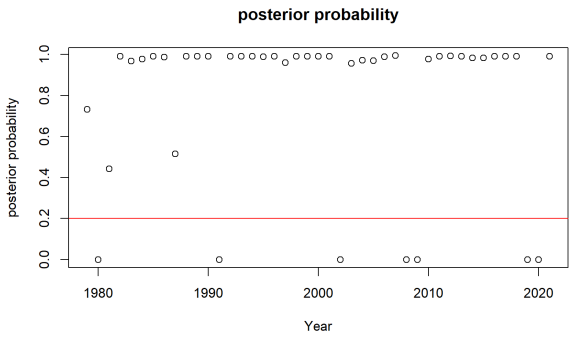

Figure 15: Posterior probability of the years after 1978

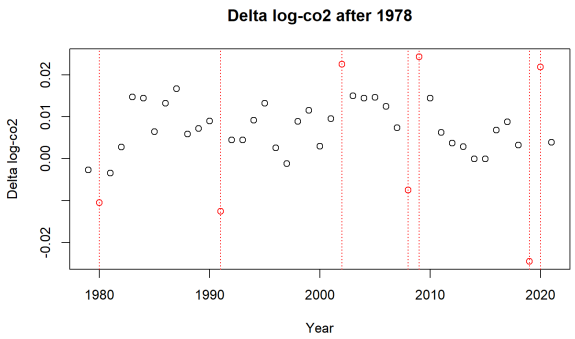

Figure 16: Delta log prediction after 1978

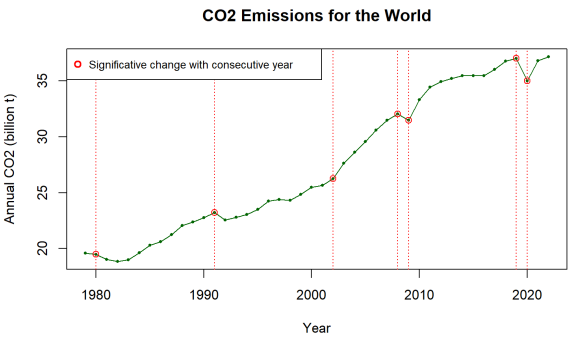

Figure 17: Prediction on real plane after 1978

3.4.4 Before 1949

We will now repeat the same analysis for the years before 1949. The general variance between datapoints is higher in this period. To obtain more extreme results we introduce a strong bias with respect to lower variance in the model and we change the priors to:

$$\begin{aligned}\beta_1 &\sim \mathcal{E}(10)\\ \beta_2 &\sim \mathcal{E}(10)\\ \delta_t &\sim Ber(p)\\ p &\sim Beta(1,1)\\ \mu &\sim \mathcal{N}(0,10)\end{aligned}$$

We choose the priors for μ, and p to be weakly informative, while for β_1 and β_2 we injected a strong bias toward values near zero (higher parameter of the exponential distribution).

3.4.5 Preprocessing

We perform the same preprocessing steps.

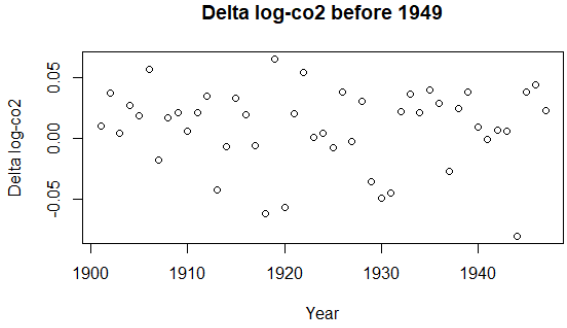

Figure 18: Data before preprocessing

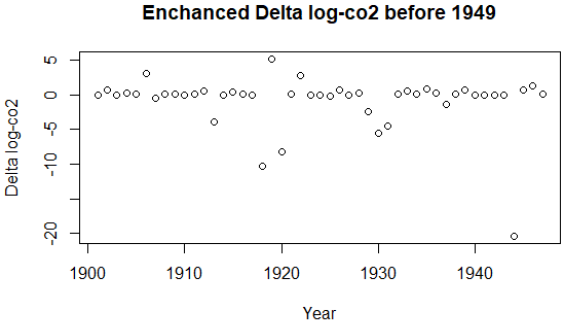

Figure 19: Data after preprocessing

3.4.6 Plots & Results

The years with the most significant variance in CO2 emission trends are: **1913**, **1918**, **1919**, **1920**, **1930**, **1931**, **1944**. Those years can be linked with the following global economic phenomena:

- 1913: **Beginning of World War I** (1914)
- 1918/1919/1920: **End of World War I** (1918) and aftermath
- 1930/1931: **The Great Depression** (1929–1939)
- 1944: **End of World War II** (1945)

Disclaimer: The years between 1900 and 1945 were troubled and a lot of significant events happened in that period. This makes the general variability between each year more accentuated and as such significant events are harder to spot.

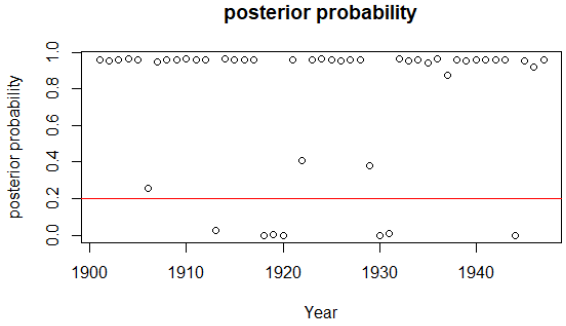

Figure 20: Posterior probability of the years before 1949

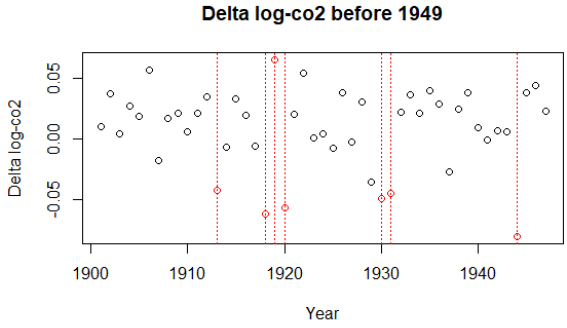

Figure 21: Delta log prediction before 1949

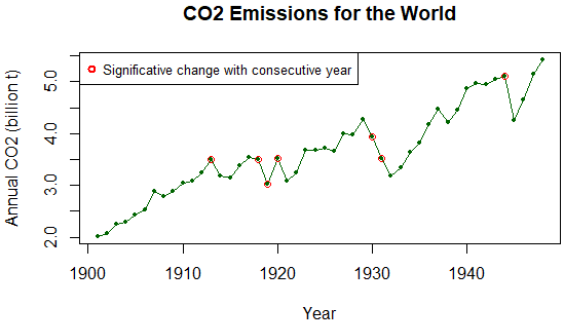

Figure 22: Prediction on real plane before 1949

4 Recent Factors

We will continue our work by analyzing the impact of various factors on CO2 emissions in recent years.

4.1 Exploratory and Correlation Analysis

We start by plotting the densities, visually representing the data in relation to CO2 emissions and analysing the correlation between covariates.

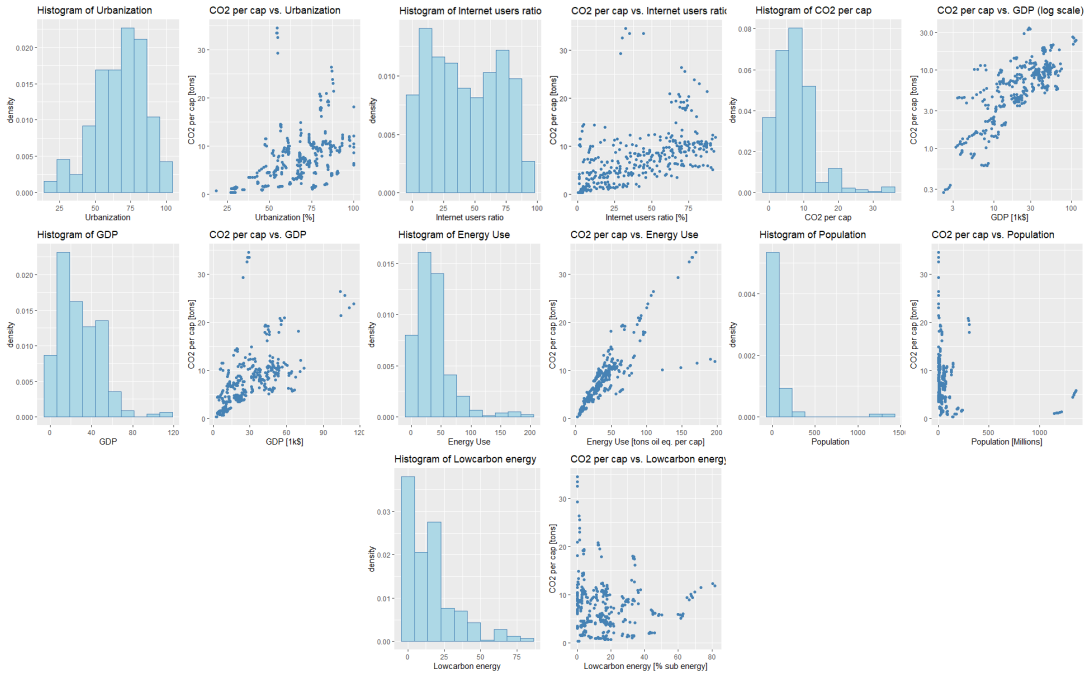

Figure 23: Density

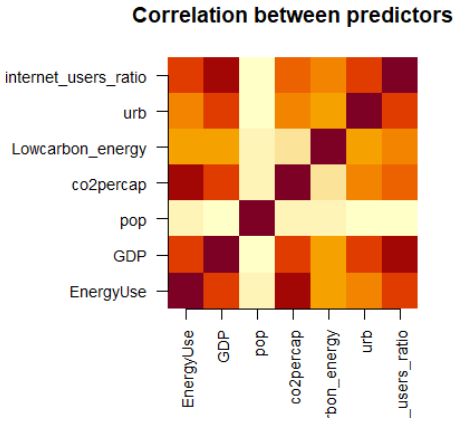

Figure 24: Correlation

As we can see the features with the highest correlation with CO2 emissions per capita (co2percap) are EnergyUse, GDP, and internet_user_ratio.

4.2 Spike and Slab prior

To determine the most impactful variables for our analysis, and perform varaible selection, we will employ a Bayesian approach using the Spike-and-Slab prior.

The spike & slab prior for a linear regression model is

$$\begin{aligned}\beta_j \mid \gamma_j &\overset{ind}{\sim} (1-\gamma_j)\,\delta_{(0)} + \gamma_j\,\mathcal{N}\left(0, \sigma^2_{\beta_j}\right),\\ \gamma_j \mid \theta_j &\overset{ind}{\sim} \mathcal{B}e\left(\theta_j\right),\\ \theta_j &\overset{iid}{\sim} p\left(\theta_j\right),\end{aligned}$$

where θ_j is a probability which determines whether β_j is nonzero and hence whether the corresponding covariate will be included in the model.

4.2.1 Posterior Probability of Inclusion

To decide whether to take in consideration a covariate we will draw a treshold on the posterior probability of inclusion.

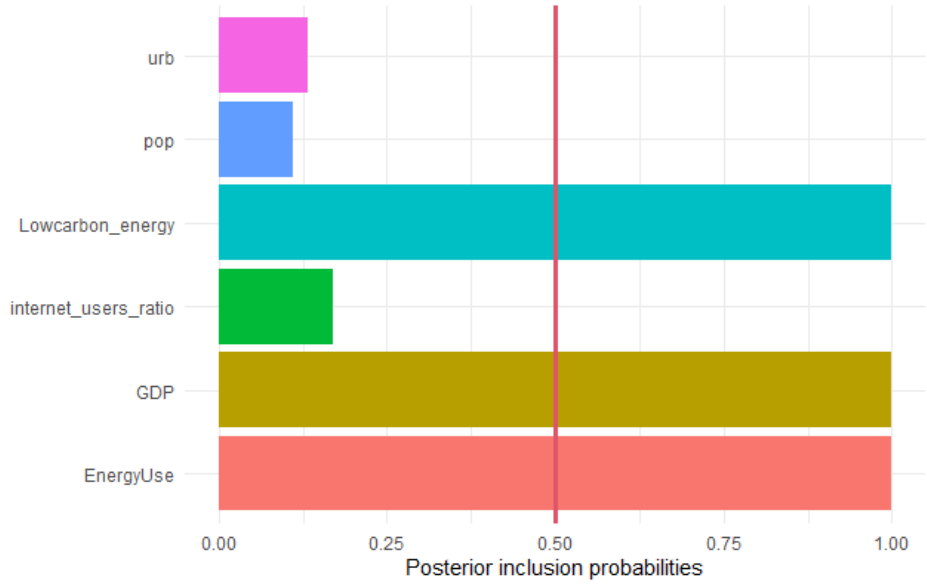

Figure 25: Posterior Probability of Inclusion

As we can see from the plot he model shows a strong preference, assigning a probability of 1 to Lowcarbon_energy, GDP, and EnergyUse, while assigning near-zero probabilities to urbanization, population, and internet_user_ratio.

These results are interesting, while the high inclusion probabilities for GDP and EnergyUse are consistent with their strong correlation with co2percap, the high probability for Lowcarbon_energy is surprising given its low correlation with co2percap.

4.3 Bayesian Regressions

From now on we will continue our analysis by focusing on the covariates GDP, EnergyUse and Lowcarbon Energy.

4.3.1 GDP and CO2 Emissions

Next, we analyze the direct relationship between CO2 emissions and GDP. We perform the regression on the log scale for better interpretability of the data.

The model is as follow:

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(\alpha + X_i\beta, \sigma^2) \\
\alpha &\sim \mathcal{N}(0, 100) \\
\beta &\sim \mathcal{N}(0, 100) \\
\frac{1}{\sigma^2} &\sim \mathcal{G}(0.001, 0.001)
\end{aligned}
$$

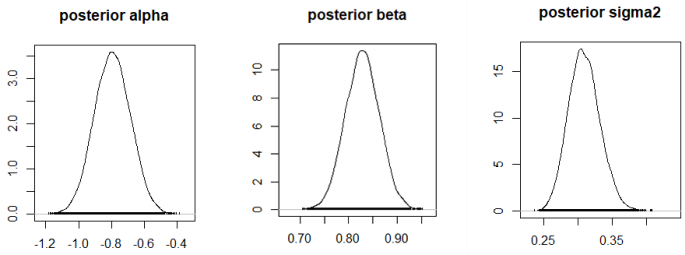

Figure 26: Posteriors

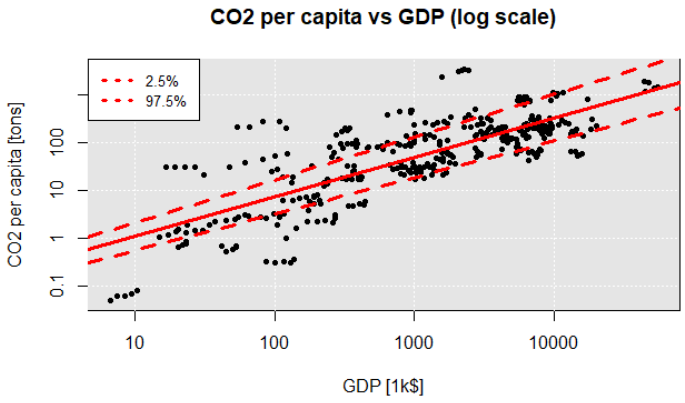

Figure 27: Linear regression over GDP

As we can see from the plot, there is a clear positive linear relationship between GDP and CO2 emissions per capita.

This finding aligns with the analysis performed on **Dataset 1**, explaining why economic **recessions** tend to reduce CO2 emission intensity, whereas periods of **economic growth** correspond with a more rapid increase in emissions.

4.3.2 Energy Use and CO2 Emissions

We also analyze the direct relationship between CO2 emissions and Energy Use. Again, we perform the regression with the same model on the log scale for better interpretability.

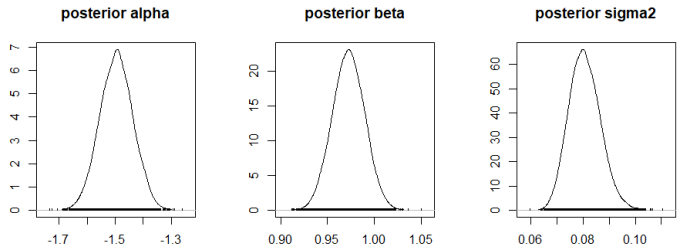

Figure 28: Posteriors

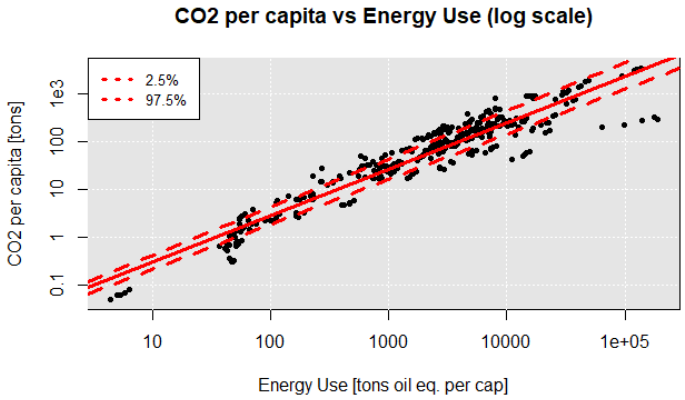

Figure 29: Linear regression over Energy use

The analysis reveals a strong linear relationship between Energy Use and CO2 emissions per capita. However, there are some data points that appear to be outliers. While these outliers are less evident in the log plot, they become more noticeable when plotted on the real scale.

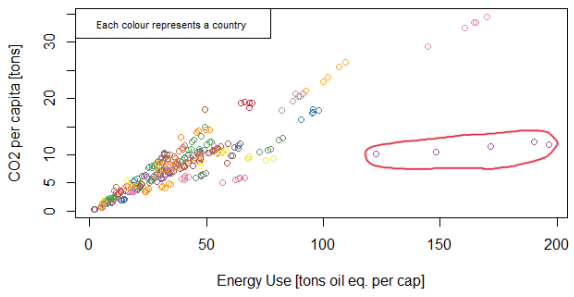

Figure 30: Linear regression over GDP

Upon closer inspection, the only outlier is Iceland. Iceland's relatively low CO2 emissions can be attributed to its substantial use of low-carbon energy sources, which accounted for up to 80% of its energy production in 2009.

4.3.3 Low carbon Energy

Low-carbon energy is defined as the sum of nuclear and renewable sources. Intuitively it is reasonable to expect to find a negative relation with the CO2 emissions. We will perform the same analysis done for GDP and Energy Use.

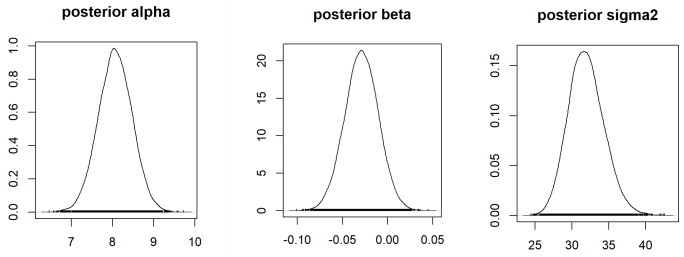

Figure 31: Posteriors

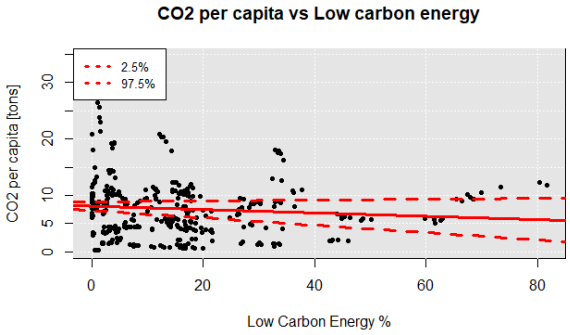

Figure 32: Linear regression over low carbon

The model we choose doesn't fit too well with the data, but a small negative trend is present. This result is not immediately evident from the plot, as the data does not appear to follow a clear negative linear trend. However it is intuitively plausible: countries with a higher proportion of Lowcarbon_energy tend to produce less CO2.

5 Conclusions & Comments

In conclusion, our analysis uncovered various insights through the application of Bayesian techniques. By analysing long term trends we found two **change points** in both variance and mean, linking them to historical events. Next we conducted a **time series** prediction exercise to see when we are likely surpass a global warming of 1.5C. And as last we fitted a **white noise model with switching variance** to our data and found correlation with major global phenomena.

After this we focused on recent factors, by employing a **spike-and-slab prior** we identified key variables impacting CO2 emissions. We then analysed them further with simple **bayesian linear regressions** to provide deeper insights.