Forecasting Renewable Energy Developments using Bayesian Time-Series Models

Lars Kutschinski

This study utilizes the World Development Indicators (WDI) database from the World Bank to forecast renewable energy usage progress over the next seven years. We analyze data spanning the past thirty years, focusing on renewable energy and other development indicators from various countries. Two distinct time-series models are employed, incorporating Bayesian methods to derive our forecasts. Results indicate significant variability in renewable energy adoption rates across countries, influenced by economic status and urbanization levels. We find that a non-zero mean AR(1) model shows good performance results. The model predicts increases in renewable energy consumption for well developed countries and decreases for countries that have more recently experienced economic growth.

Introduction

In recent years, the shift toward renewable energy has been recognized as crucial in the global effort to mitigate climate change, reduce greenhouse gas emissions, and enhance sustainable energy security. The pace at which renewable energy is adopted and integrated into national grids varies significantly across countries, driven by factors that include economic conditions and urban development. This project focuses on understanding how these specific elements—GDP per capita and urban population percentage—affect the adoption of renewable energy. Furthermore we will focus on making forecasts for the next seven years of renewable energy development. An argument could be made that a more enticing problem of interest would be forecasting the production of renewable energy instead of consumption, but we argue that these problems are more or less the same, since production of energy goes hand in hand with it's consumption.

Numerous global organizations, such as the World Bank, are instrumental in advancing sustainable development and in tracking the progress of such initiatives through the aggregation and analysis of data. A valuable asset in these efforts is the World Development Indicators (WDI) database, managed by the World Bank, which provides comprehensive data on various development indicators spanning several decades. Utilizing this rich dataset, our analysis will employ time series models to examine the influences of GDP per capita and urban population percentage on renewable energy consumption. Furthermore, we aim to forecast future trends in renewable energy usage, which is essential for crafting effective policies and interventions to accelerate its adoption in alignment with global sustainability objectives. This timely exploration offers insights into the dynamics shaping renewable energy growth, highlighting the importance of economic and urban factors in shaping sustainable energy landscapes.

Data

As previously mentioned, our analysis will utilize data from the World Development Indicators (WDI) database maintained by the World Bank. This extensive dataset contains 1,486 development indicators, collected from a variety of officially recognized international sources spanning the years 1960 to 2022. While the WDI includes data for 266 countries, our study narrows the focus to a selected group of 10 countries to manage the complexity of our analysis. This selection represents a diverse array of development stages: highly developed nations such as the United States, Canada, and Germany; emerging economies like China and India; countries heavily reliant on renewable energy including Iceland and Sweden; and developing nations such as Brazil and Uganda. This strategic choice of countries is designed to effectively capture the varying trends in renewable energy consumption across different economic landscapes

Countries
Brazil
Canada
China
France
Germany
Iceland
India
Sweden
Uganda
United States

Now we will briefly go over the variables that are present in our data and that we will use for the regression analysis:

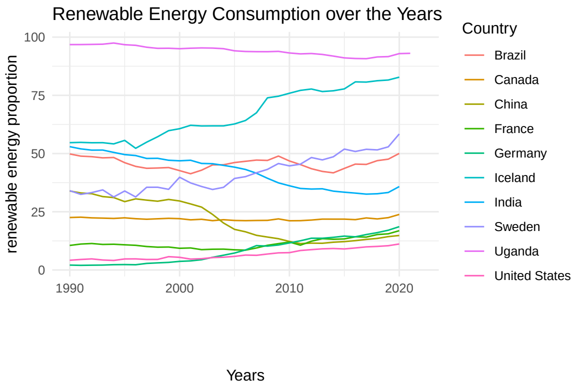
- Renewable energy (renewable_energy): The proportion of renewable energy consumption in total final energy consumption.
- Country (country): Selected countries
- Years (year): Time period of data collection (1960-2022)
- GDP per capita in US$ (gdp_per_capita): Measures the average economic output per person

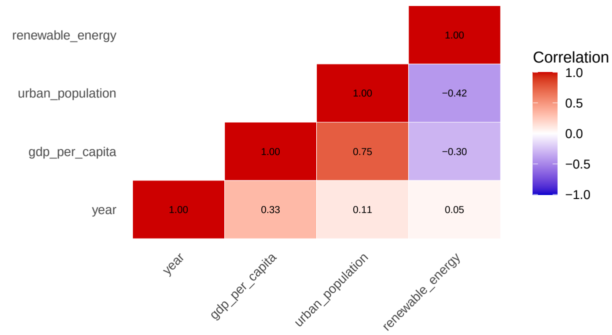
country	year	gdp_per_capita	urban_population	renewable_energy
Brazil	1960	NA	46.139	NA
Brazil	1961	NA	47.122	NA
Brazil	1962	NA	48.099	NA
Brazil	1963	NA	49.078	NA
Brazil	1964	NA	50.059	NA
Brazil	1965	NA	51.037	NA

country	year	gdp_per_capita	urban_population	renewable_energy
Brazil	1990	3065.242	73.922	49.82
Brazil	1991	2656.497	74.690	48.89
Brazil	1992	2505.378	75.444	48.65
Brazil	1993	2766.346	76.181	48.15
Brazil	1994	3393.143	76.903	48.30
Brazil	1995	4704.960	77.610	46.08

- Urban population as % of total population (urban_population): Indicates the proportion of the country's population living in urban areas

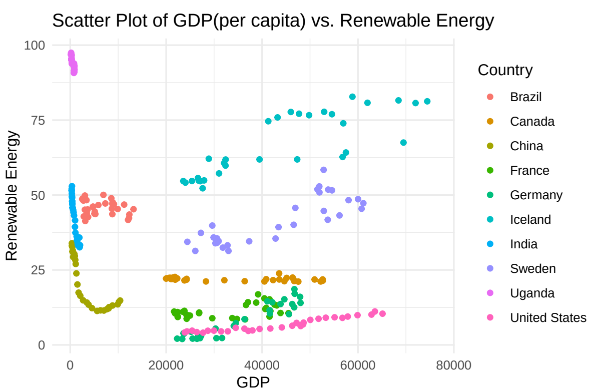
Missing Data: The World Development Indicators (WDI) database represents a monumental effort to amass a comprehensive array of global development data. Given the ambitious scope of this endeavor, it's understandable that gathering exhaustive data for every indicator across numerous countries and over an extended timeline presents significant challenges. Certain nations, particularly those with more restrictive data-sharing policies, may contribute to gaps in the dataset, leading to a notable presence of missing data entries. Notably, data on renewable energy consumption and GDP per capita within the WDI only extends back to 1990, reflecting the evolving focus of global development metrics. In our analysis, we will adapt to this limitation by excluding data from before 1990, aligning our study period with the availability of reliable renewable energy statistics. Furthermore, the data mostly only contains renewable energy information up to 2020. Since we are making predictions for the next 7 years these prediction will be from 2021 to 2027.

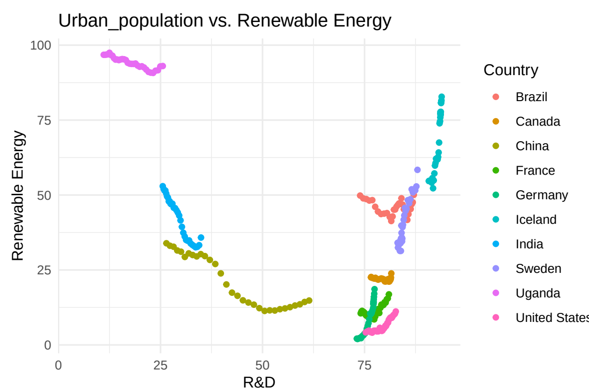




In the left figure we illustrated the change in renewable energy consumption in the last 30 years. We see that a lot of countries, most notably Sweden have made good progress in increasing the consumption proportion. On the other hand, India and China's renewable energy consumption rate has actually decreased significantly, which is likely due to the increasing economic growth that those countries are investing in.

In the right figure we have gathered the correlations of the variables of our data. A first trend that we notice is that gdp and urban population percentages seem to be negatively correlated with renewable energy consumption. This however could be a misleading conclusion in some cases: In the left plot we see that countries like Uganda and Iceland lead the renewable energy consumption while having comparatively low gdp's. Iceland is unique country that is rich in natural resources that make renewable energy production easier than anywhere else in the world.





These two plots confirm that, while the covariates and the response appear negatively correlated in some countries, other countries show that an increase in GDP and urban population leads to an increase in renewable energy production.

Methods

We will be making use of two different time-series models for our analysis. We have data ranging from 1990 to 2020 and expect there two be an auto-regressive relationship in the response variable renewable_energy. Hence AR time process modelling makes sense here.

Model 1

The first model that we use is a standard non-zero mean AR(1) model. The reason for the zero-mean addition is that we model the response variable at time t using the information at time t-1 of the response as well as our two covariates gdp_per_capita and urban_population. The model thus takes the following form:

$$y|\gamma_t, \sigma \sim N(\gamma_t, \sigma^2)$$

$$\gamma_t = \beta_{\text{gdp}} * gdp + \beta_{\text{urban}} * urban + \mu_t, \text{ with } \mu_t \sim AR(1)$$

The AR(1) process μ_t is defined as:

$$\mu_t = \rho * y_{t-1} + \varepsilon_t$$

$$\varepsilon_t|\sigma \sim N(0, \sigma^2)$$

Model 2

The second model is a second order non-zero mean Random Walk model. This model is defined as:

$$y|\gamma_t, \sigma \sim N(\gamma_t, \sigma^2)$$

$$\gamma_t = \beta_{\text{gdp}} * gdp + \beta_{\text{urban}} * urban + \mu_t, \text{ with } \mu_t \sim AR(2)$$

The AR(2) process μ_t is defined as:

$$\mu_t = \rho_1 * y_{t-1} + \rho_2 * y_{t-2} + \varepsilon_t$$

$$\varepsilon_t | \sigma \sim N(0, \sigma^2)$$

In order to have a stationary distribution the following conditions of ρ_1 and ρ_2 are required:

- $\rho_1 + \rho_2 < 1$
- $\rho_1 - \rho_2 < 1$
- $|\rho_1| < 1$

Priors

Both models utilize mostly the same priors. We will define uninformed priors for each of our parameters:

$$\beta_{\text{gdp}} \sim N(0, 1)$$

$$\beta_{\text{urban}} \sim N(0, 1)$$

$$\sigma \sim N^+(0, 1)$$

$$\rho \sim Unif(-1, 1)$$

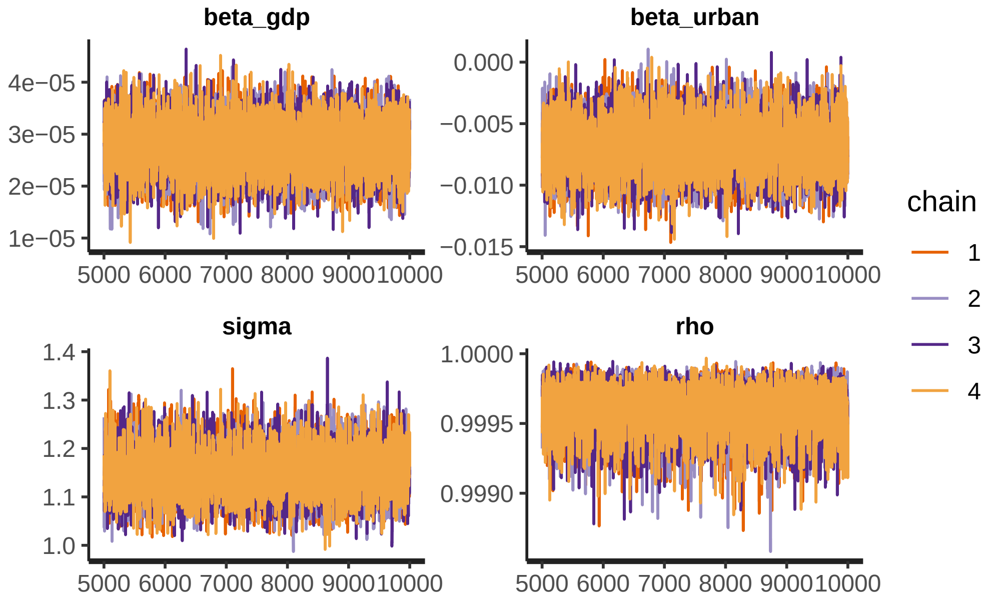
$$\rho_1 \sim Unif(-1/2, 1/2) \text{ and } \rho_2 \sim Unif(-1/2, 1/2)$$

Results

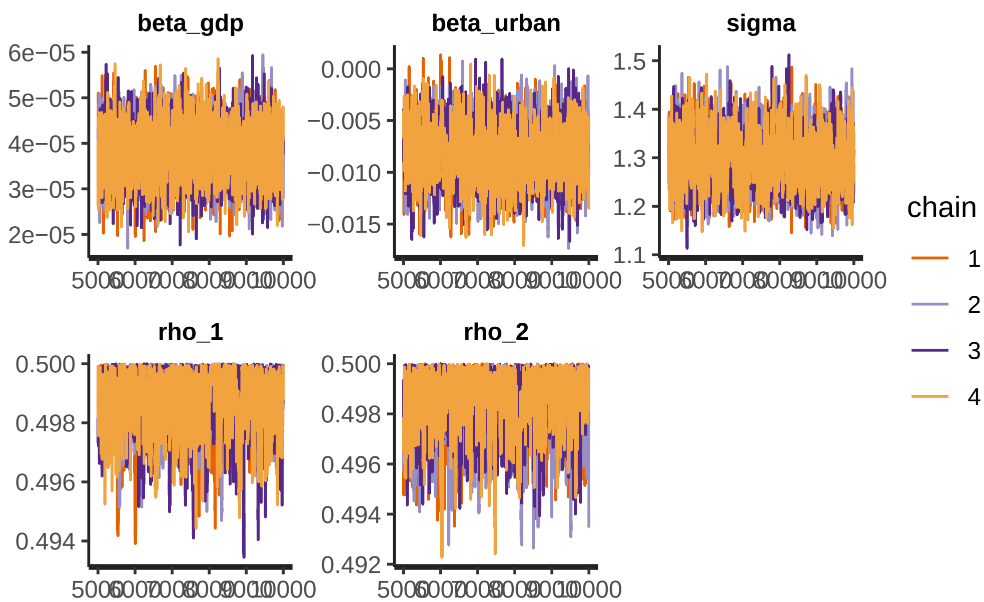
Model Convergence and Validation

We ran both of our models in rStan using the indicated priors for the parameters. The number of iteration is set to 10000, the number of Markov chains is set to 4 and the seed is set to 3105. The code for our models can be found in the repository https://github.com/LarsKut/STA2024/tree/main/Final%20Project.

Traceplot for Model 1



Traceplot for Model 2



First we analyze the traceplots for model 1 and 2. In both models we notice good mixing of the MCMC chains for all parameters.

Summary of Model 1

	mean	se_mean	sd	2.5%	25%
beta_gdp	2.767232e-05	5.487095e-08	4.607743e-06	1.862782e-05	0.0000245743

```
beta_urban -6.804298e-03 2.640184e-05 2.045253e-03 -1.081390e-02 -0.0081876282
sigma       1.152021e+00 7.155660e-04 4.773485e-02  1.063614e+00  1.1187322843
rho         9.996177e-01 2.250707e-06 1.586267e-04  9.992485e-01  0.9995273108
                     50%           75%         97.5%    n_eff     Rhat
beta_gdp    2.768629e-05  3.073817e-05  3.674859e-05 7051.664 1.000339
beta_urban -6.806299e-03 -5.424509e-03 -2.784048e-03 6001.030 1.000562
sigma       1.150482e+00  1.183415e+00  1.249154e+00 4450.121 1.000509
rho         9.996402e-01  9.997330e-01  9.998598e-01 4967.235 1.000212
```

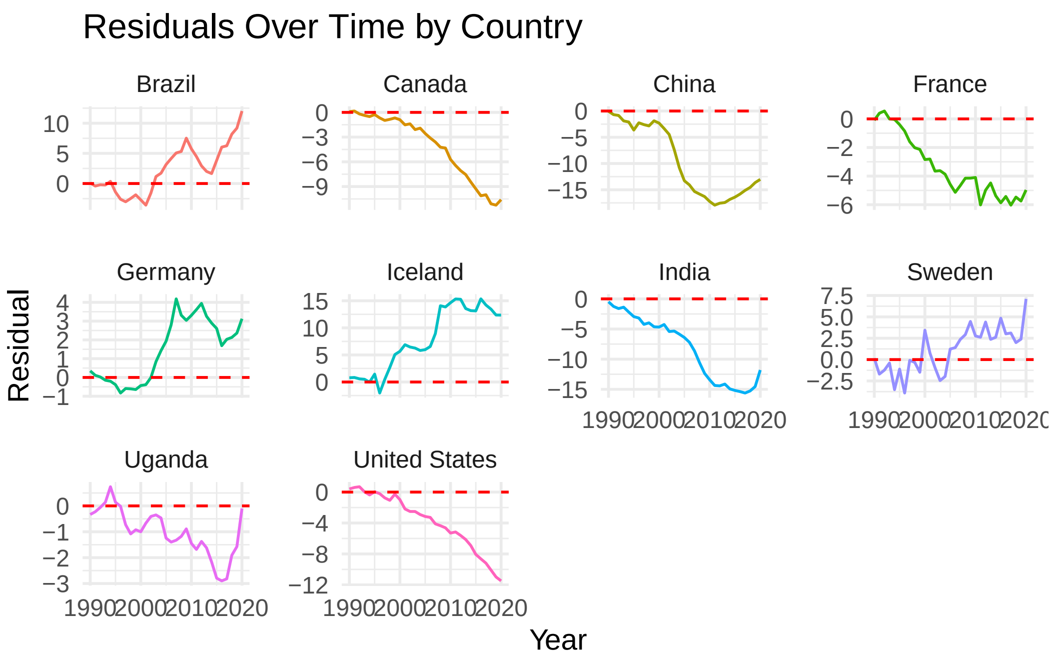
Summary of Model 2

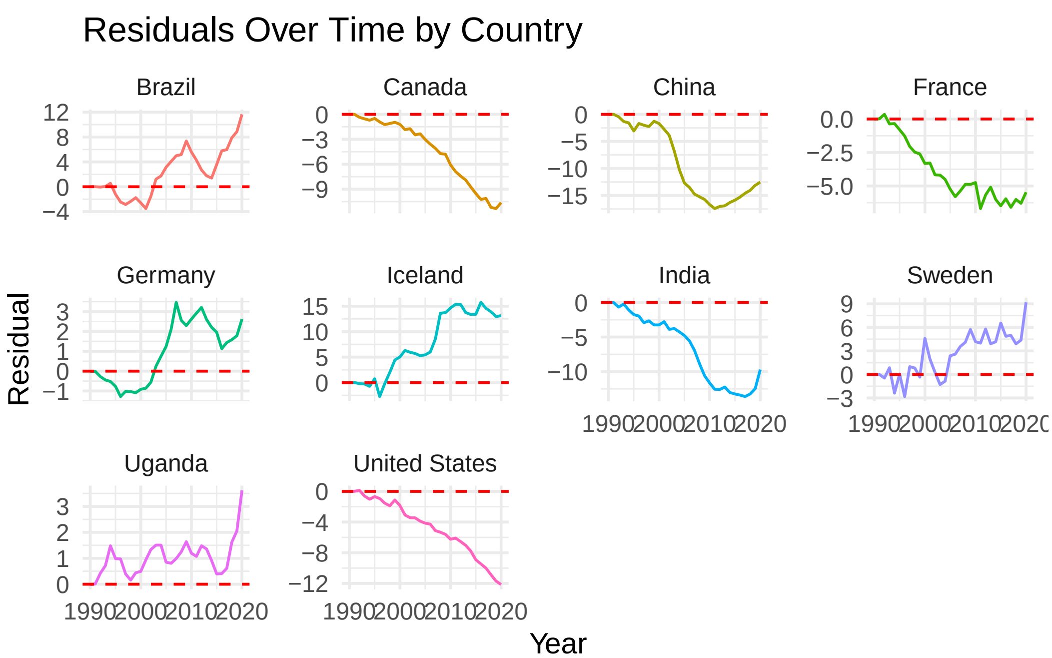
```
                    mean      se_mean           sd          2.5%           25%
beta_gdp    3.843038e-05 1.240811e-07 5.350852e-06  2.803464e-05  3.484768e-05
beta_urban -8.404195e-03 7.049246e-05 2.550045e-03 -1.355301e-02 -1.007446e-02
sigma       1.296558e+00 1.286610e-03 5.100707e-02  1.201862e+00  1.261332e+00
rho_1       4.987610e-01 3.904018e-05 1.023055e-03  4.962078e-01  4.982083e-01
rho_2       4.984957e-01 4.078994e-05 1.216978e-03  4.953609e-01  4.978820e-01
                     50%           75%         97.5%     n_eff     Rhat
beta_gdp    3.837518e-05  4.193804e-05  4.917862e-05 1859.6650 1.001622
beta_urban -8.395672e-03 -6.720126e-03 -3.359420e-03 1308.6100 1.003039
sigma       1.294589e+00  1.329804e+00  1.402392e+00 1571.6910 1.000666
rho_1       4.989897e-01  4.995711e-01  4.999594e-01  686.7113 1.002583
rho_2       4.987784e-01  4.994336e-01  4.999426e-01  890.1419 1.006006
```

In the summary plots for our model we also notice Rhat values below 1.05 and high effective sample sizes for all parameters, further indicating good convergence results. We also observe that both models estimate very small effects for the coefficients of the covariates.

Model Validation

Model Validation for time-series can be a bit difficult, since regular posterior predictive check methods do not work as well. The main indication for a good fit that we will use is Residual plots. We plotted the residuals across all years between 1990 and 2020 for both our models:

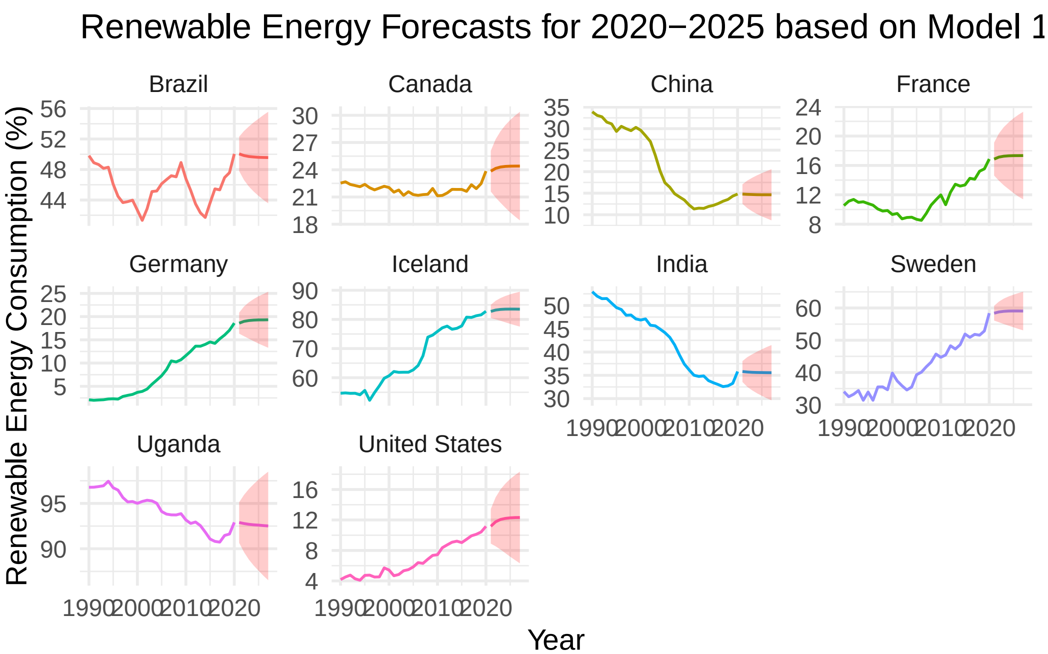


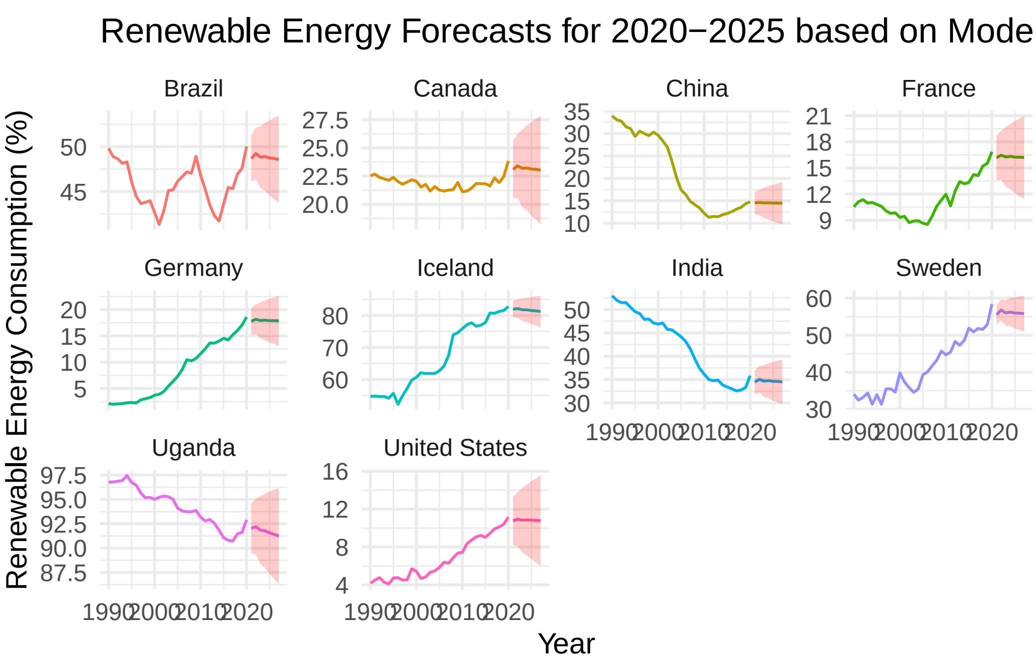


We can see that the residual plots of both models are nearly exactly the same. After checking the posterior means of the response we observe that indeed the difference in residuals is nearly zero between the two models. Hence the conclusion is that model 1 could be better than model 2, since it is a simpler model.

Forecasting Results

We will now compare the forecasting results of our two models. We have forecasted the renewable energy consumption over the next seven years for each of the selected countries:





We observe that, while the forecasts are also quite similar, the AR(2) model seems to predict more of a decrease in renewable energy consumption compared to model 1. Model 1 suggests that renewable energy consumption will increase in the next few years in countries such as Germany, France, Canada, Sweden and the US. On the other hand it predicts that consumption will decrease again in China, India and Brazil.

Discussion and Conclusion

We have developed two forecasting models for the renewable energy consumption in ten different countries. It was found that the AR(1) model should be preferred for this forecasting problem. The model predicts that the renewable energy usage will increase in well developed countries such as Germany, France, Canada, Sweden and the US, but will decrease again in countries such as China, India and Brazil. The conclusion is quite interesting, since it suggests that countries that have enjoyed a good economical status for a long time (indicated by GDP for example) are able to increase their renewable energy production while maintaining high economical wealth. Recent strong economic developments in countries like China have relied partly on non-renewable energy investments, posing challenges to increasing clean energy usage without economic sacrifices.

This analysis can be extend to be performed for multiple more countries and is quite useful therefore. It can be used to forecast the progress towards renewable energy and to determine countries that currently pose a barrier towards the progress. Then institutions such as the World Bank can look into measures of helping such countries get on the right track.

So far we have only used two important indicators as covariates for this time series model. Future work could look into adding more covariates that could potentially have high correlations with renewable energy. Other time series models might also be worth looking into as well. For model comparisons the LOO-LFO validation could be implemented, which is based on the classical LOO validation of bayesian models. This technique is especially crucial in time series analysis due to the sequential nature of the data, where conventional random splitting used in cross-validation can disrupt the temporal order, leading to unrealistic training and testing scenarios.

References

Dataset: Dataset Link

Github Repository: https://github.com/LarsKut/STA2024/tree/main/Final%20Project