Climate Change Analysis in the United States: The Impact of Socioeconomic and Industry on Average Temperature and Wolf Sunspot Cycle Spectral Analysis

Sasha Matveev netID(matveev2), Vladislav Fedorov netID(vvf2)

Author Note

The authors made the following contributions. Sasha Matveev worked on Analysis A including all code, analysis, preparation, and writing. Vladislav Fedorov worked on Analysis C including all code, analysis, preparation, and writing. Together the authors worked on formatting the report and editing the content.

Abstract

The goal of the study is split into two parts - the goal of Analysis A is to investigate the relationship between social and industrial factors contributing to climate change. We measure this using the average annual temperature with a focus on the United States from 1900 to 2023, aiming to determine strongest contributors and predictors of the average annual temperature. The goal of Analysis C is to investigate and confirm the underlying periodicity of Wolf sunspot solar cycle using the data containing the mean number of sunspots across approximately 300 years. In Analysis A we compared three models fit exclusively on social factors, exclusively on emissions factors, and a third model fit on both categories of climate change factors to determine the strongest predictors of the average annual temperature. We found that the emissions model provided the best fit for the data and was able to forecast future values, capturing actual average annual temperatures within the 95% confidence intervals for 2023 and 2024. The results of the study underline the complexity of climate change while also showing that statistical models can be used to inform lawmakers in potential mitigation efforts. In Analysis C we used multiple spectral analysis tools, which included periodograms, wavelet analysis, and multitaper analysis, and determined that the most dominant period of the sunspot solar cycle is approximately 11 years. The result is consistent with the past research, confirming the effectiveness of current methodology to model solar cycle periodicity.

Analysis A: Climate Change

Introduction

Climate change is listed as one of the most critical challenges facing humanity today by the United Nations, affecting everything from "…shifting weather patterns that threaten food production, to rising sea levels that increase the risk of catastrophic flooding, the impacts of climate change are global in scope and unprecedented in scale. Without drastic action today, adapting to these impacts in the future will be more difficult and costly"(United Nations). There are numerous social, industrial, geographic, economic, and historic factors that contribute to climate change, making it a complex issue that is difficult to model. As such, over the course of this study we will focus primarily on the industrial factors such as greenhouse gas emissions and socio-economic factors such as changes in population and gross domestic product (GDP). This study aims to analyze the social and environmental factors that contribute the most to climate change in the United States at the individual level as well as in a combined setting, withe the impact tracked by the average annual temperature in the country from 1900 to 2022. Through regression modeling and time series analysis, we look to identify the strongest contributors to climate change as well as the factors that offer the best predictive power for forecasting future temperature changes in the US so that policy makers can better understand the issue and have reliable data to base potential mitigation strategies on.

Data Overview

The original emissions data comes from Our World in Data and was compiled by Shreyansh Dangi for use on kaggle. The data includes global carbon dioxide (CO2) and other greenhouse gas emissions across countries, and sectors while also combining variables describing the countries in the dataset such as population, GDP, and energy consumption. Each row is a country and a given year so the time unit is yearly. The analysis in this study will be focusing on the United States from 1900-2022. The variables of interest are time (year), the population, GDP (inflation adjusted and cost-of-living adjusted in dollars), CO2 emissions (million tons), methane emissions (tons), nitrous oxide emissions (tons), as well as the average annual temperature in Fahrenheit from the supplementary dataset. The supplementary dataset containing the average annual temperature in the US comes from the National Centers for Environmental Information (NCEI) and was made available on kaggle by Gia Bách Nguyen. The data is based on observations from a network of

thousands of weather stations across the United States in the same exact time period (1900-2022) as the primary dataset containing emissions data and is also yearly.

Statistical Methods:

Preliminary analysis

The temperature time series was analyzed through exploratory data analysis techniques, including both statistical and visual inspections of graphs. The time series plot in Figure 1 does not show any seasonality but does show a very weak upward trend over time, as expected based on prior background knowledge of rising temperatures.

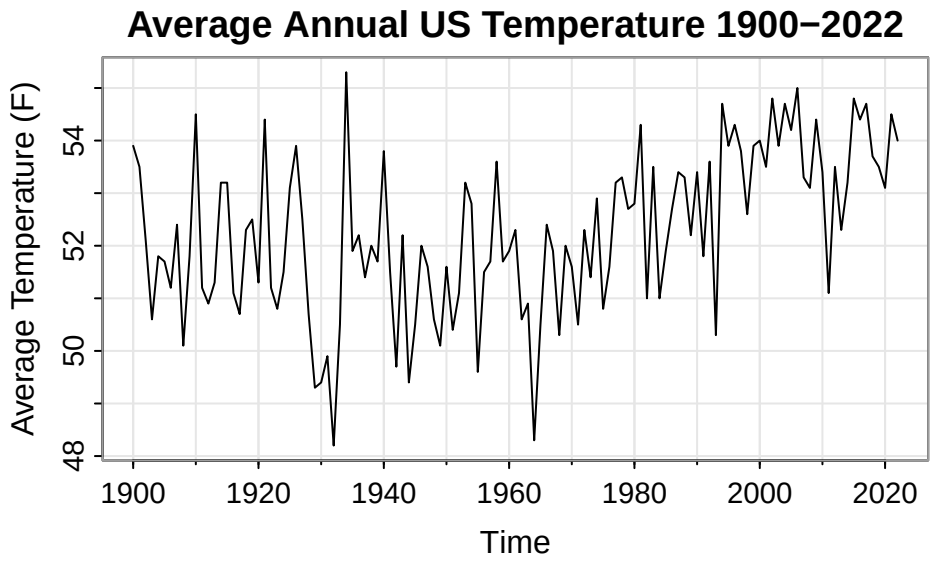


Figure 1: Time Series Plot of Average Annual US Temperature (1900-2022)

The variance plot does not appear to show a pattern of either increasing or decreasing over time. Furthermore, to examine the variance through a statistical manner, a Levene's test was performed by separating the data into two groups chronologically up to the halfway point. The result was a p-value of 0.64 so there was not enough statistically significant evidence to conclude a difference of variance between our two groups, giving more credability to the claim of having a lack of change in variance over time. The variance plot can be seen in Figure 2 below.

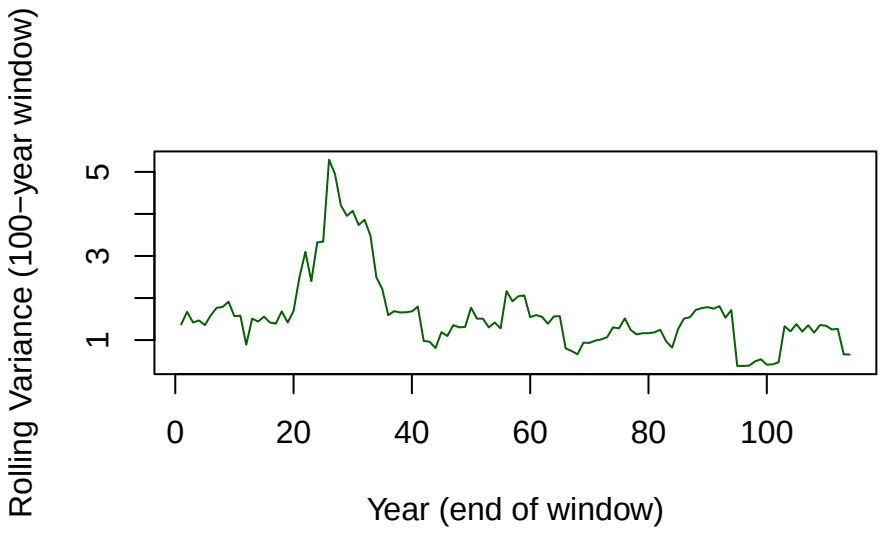


Figure 2: Rolling 100-Year Variance of US Temperature

Based on the results of the Augmented Dickey-Fuller test we have a p-value = 0.0224 so we have enough statistically significant evidence to reject the null hypothesis of non-stationarity and conclude that the temperature time series is actually stationary. Thus we do not need to perform any further transformations to the data to achieve stationarity prior to fitting regression models.

Regression Models

In order to examine the questions posed by the study we will fit three regression models to predict the average annual temperature in the US. The models will differ only in the variables they are fitted on, with the first model will include all potential predictors, both emissions and social including : time (year), population, GDP, CO2 emissions, methane emissions, and nitrous oxide emissions. The second model aims to test the impact of social and demographic factors on the average annual temperature in the US. Therefore the second model will include only social predictors: time (year), population, and GDP. Lastly, the third model is constructed to isolate the emissions factors and examine the relationship between time (year) and the three most common types of emissions including CO2 emissions, methane emissions, nitrous oxide emissions with the average annual temperature in the US between 1900 and 2022.A succinct summary of the variables in the three models is shown in Table 1 below.

Table 1: Summary of Variables in Regression Models

Model	Predictors
Full Model	year, population, gdp, co2, methane, nitrous_oxide
Social Factors Only	year, population, gdp
Emissions Only	year, co2, methane, nitrous_oxide

The reason behind the categories of variables chosen is to test the impact of varying climate change factors both independently and together. The full model will allow us to see the combined effect of both social and emissions factors on the average annual temperature in the US. The social factors only model will help us understand how much of the change in temperature can be explained by demographic and economic factors alone, while the emissions only model will isolate the effect of greenhouse gas emissions on temperature

changes, which is often attributed as the driving force behind climate change in the time period examined in this study.

Regression Result Analysis

The full model including all predictors, both social and emissions factors, showed that only the CO2 predictor was significant at alpha =0.05 level when performing the t-test. No other predictor came even close to being significant. The adjusted R-squared value for the full model was 0.3272 , indicating that approximately 32.72 % of the variance in average annual temperature in the U.S. between 1900 and 2022 can be explained by the proposed model. Full model summary and diagnostics for each model can be found in the Appendix section.

The social factors only model showed that none of the predictors were significant at alpha =0.05 level. The closest variable to being significant was gdp with a p-value of 0.0599. Like in the full model, the year and population predictors were not significant. The adjusted R-squared value for the social factors only model was relatively high at 0.3147, meaning 31.47% of the variance in the average U.S. annual temperature between 1900 and 2022 can be explained by the social factors only model, where no predictor was significant at the alpha=0.05 level.

The emissions model had two predictors that were significant at the alpha=0.05 level, including CO2 and methane emissions with p-values of 0.000531 and 0.000520, respectively. The nitrous oxide predictor was not significant with a p-value of 0.120. The adjusted R-squared value for the emissions only model was 0.3314 , indicating that 34.13% of the variance in average annual temperature in the U.S. between 1900 and 2022 can be explained by the emissions only model. The year variable was far from being significant once again with a p-value of 0.5956 but the nitrous oxide variable was much closer to being significant.

The model diagnostics for all models are relatively similar despite having different predictors. The residuals vs fitted values plot shows no signs of nonlinearity nor heteroscedasticity for all models. The Q-Q plots for all models show slight deviations from normality in the tails but given the relatively small dataset size of 123 observations this is expected. Observation 35 was noted to be a potential point to be removed due to moderate influence however based on domain knowledge we decided to keep it in the dataset as it was characteristic of the wild weather patterns of the 1930s which saw devastingly cold winters and the infamous Dust Bowl which wreaked havoc across the Midwest.

Regression with Autocorrelated Errors

Since the residuals of all models showed no signs of autocorrelation based on the acf plot and the Ljung-Box test p-values were all higher than alpha =0.05 , there was no need to fit regression models with autocorrelated errors.

Final model selection

The lowest AIC value was observed in the emissions only model with an AIC = 411.5568. The lowest BIC value belongs to the social model with a BIC of 427.6950. Lastly, the highest adjusted R^2 value belongs to the emissions only model as mentioned previously with a value of .3314. A complete breakdown of the AIC, BIC and adjusted R-squared values for all models is shown in Table 2 below.

Table 2: Model Comparison Metrics

Model	AIC	BIC	Adjusted R-squared
Full Model	414.2226	436.7201	0.3272
Social Factors Only	413.6341	427.6950	0.3147
Emissions Only	411.5568	428.4299	0.3314

We know that lower AIC and BIC values indicate a better fitting model while higher adjusted R-squared values indicate a better fitting model. Furthermore, BIC tends to prefer smaller models. Based on the model comparison metrics, the emissions only model will be selected as the final model for forecasting purposes since it has both the lowest AIC value and the highest adjusted R-squared value while having a BIC value very close to the lowest one observed in the social model

Forecast future 5 values

The best model found was the emissions only model which included year, co2, methane, and nitrous oxide as predictors. Using this model we forecasted the average annual temperature in the US for the next 5 years (2023-2027) based on a growth rate of 1 percent per year for each emissions predictor. The forecasted values are shown in Table 3 below.

Table 3: Forecasted Average Annual US Temperature (2023–2027)

Year	Forecasted Temperature (F)	Lower 95% CI	Upper 95% CI
2023	53.4862	50.8918	56.0806
2024	53.5161	50.9212	56.1109
2025	53.5463	50.9511	56.1416
2026	53.5771	50.9814	56.1727
2027	53.6082	51.0121	56.2043

Since the official temperature data for 2023 and 2024 has been released by NCEI, we can compare our forecasted values to the actual observed values. In 2023 the actual average annual temperature in the US was 54.4 F while our forecasted value was 53.48625 F. In 2024 the actual average annual temperature in the US was 55.5 F while our forecasted value was 53.51607 F. Thus our model underestimated the average annual temperature in both years by approximately 0.9 F in 2023 and 1.98 F in 2024. This discrepancy is not surprising given the simplicity of our model and the fact that we assumed a constant growth rate of 1 percent per year for each emissions predictor. It should be noted however, that the point estimate is oftentimes rarely accurate in predicting future values, especially as we extrapolate further from the observations used to fit the model. In looking at the confidence intervals for our forecasted values, we can see that both actual observed values for 2023 and 2024 fall well within the 95% confidence intervals of our forecasts for the average annual temperature, indicating that while the point estimates were not accurate, the model was still able to forecast future values moderately well.

Discussion & Conclusion

Overall, the best model proposed by this study explains only 33.14% of the variance in average annual temperature in the US between 1900 and 2022. The result is indicative of the complexity of the issue of climate change as a whole, because there are numerous factors that contribute to it beyond just the three main greenhouse gas emissions explored by the study within the emissions model. Given that the year variable was not significant in the final model, the findings suggest that the increases in average annual temperature in the US over time are not driven by time components but rather by external socioeconomic factors among which are greenhouse gas emissions. This finding aligns with the stationary of the temperature time series observed in the preliminary analysis based on the adf test results. The two most significant predictors based on the lowest p-values from the t-test in the emissions model were somewhat unsurprisingly CO2 and methane emissions as they are widely regarded to be among the most common greenhouse gasses contributing to climate change.

Another potential major shortcoming of this model is its simplicity as we include a relatively small number of predictors (four) with a total of 123 observations from just a single country. The climate and socioeconomic

patterns observed within the United States are not entirely representative of global trends so the model may struggle to generalize to other countries or continents. The decision to keep the models small was mainly done for the purpose of preserving interpretability while also keeping the task manageable from a time perspective.

Future studies may explore various avenues of improving upon the results achieved by this exploration. As mentioned previously, expanding the dataset to include more countries and potentially more variables such as deforestation rates, industrial activity levels, energy production and consumption could all assist in making the model more robust and capture more of the variance in the average annual temperature. Making the model more robus may also increase the accuracy of its forecasting although more complex models are more likely to overfit so a balance should be struck. In addition, a different target variable may be used to model different aspects of climate change such as sea level rise. Furthermore, a potential area of improvement to the current modeling approach that is relatively easy to implement but was beyond the scope of the study is the use of model selection techniques such as forward, backward, or stepwise selection to remove statistically insignificant predictors from the results.

The study shows that the primary factors associated with a change in the average annual temperature in the U.S. between 1900 and 2022 were CO2 and methane emissions. Another factor that showed some association with the average annual temperature was GDP in the social factors only model which makes sense as GDP is a measure of economic activity and industry which produce greenhouse gas emissions. The emissions model fit with the CO2 and methane emissions predictors offered moderate forecasting power for the average annual temperature over the next five years with confidence intervals that contained the actual observed values for 2023 and 2024. The techniques implemented in this study not only highlight the complexity of climate change as an issue but also highlight the potential to model and forecast its effects using statistical methods to provide lawmakers with actionable insights that could be used to alleviate its impact on humanity in the coming century.

Analysis C: Sunspot Spectral Analysis

Introduction

Magnetic activity of the Sun is a crucial topic in astronomy, both today and in the past, because it is a driving factor in solar flares and solar storms which can affect function of satellites, power grids and communication systems on Earth. Knowing when to expect flares and storms can help in preparing for these potentially disruptive and dangerous events so we can mitigate their consequences with minimal damage. The number of sunspots tracked across the years serves as a visible record of this magnetic activity, letting us see when the Sun is entering more energetic phases. Using spectral analysis, we will study the mean number of sunspots per year across more than three centuries worth of data and examine whether the underlying period of the solar cycle is consistent with previous domain knowledge or whether it has shifted over time.

Data Overview

The data for this analysis comes from the Royal Observatory of Belgium. It contains the following variables of interest: year - calendar years ranging from 1700.5 to 2024.5; mean number of sunspots - the average number of sunspots detected on the Sun's surface during the corresponding calendar year. Analyzing variations the average number of sunspots each year will provide an insight into the solar cycle's periodicity.

Literature Review

Spectral analysis is used to detect cycles in a time series by examining its frequency structure. Basic tools like the periodogram can reveal strong cycles, but they are often quite noisy, so methods like the multitaper or wavelet approaches are used to get smoother and more reliable results. In a related paper, Berger A., et al (1990) also discuss amplitude and phase estimation using, for instance, complex demodulation. These techniques are especially helpful when the underlying cycle is not perfectly steady over time. For example, in the sunspot data, the length of the solar cycle drift across decades, making it a good case for more advanced spectral tools like wavelet transform and miltitaper analysis. With several centuries of observations, the dataset clearly shows both the dominant 11-year cycle and how its behavior varies over time.

Statistical Methods:

Basic Time Series Analysis

Since this is a time series analysis first and foremost, the initial visualization we wanted to see is the basic time series plot, which is shown below.

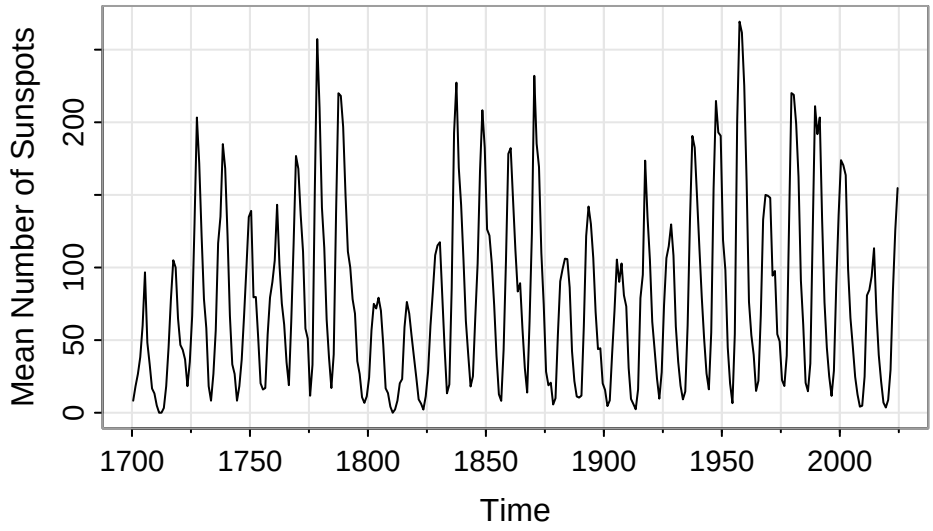


Figure 3: Time Series Plot for Sunspot Data

From the plot, the data clearly has a strong sinusoidal structure. The amplitude varies over time, and the cycle length does as well, indicating that the oscillation is not perfectly regular. With that said, the ADF test (the code for which can be found in the Analysis C section of the Appendix) suggests the series is stationary, although this does not fully affect the choice of methods here, and it is simply an interesting observation.

Periodograms

When it comes to spectral analysis, the most common first approach is a raw periodogram - a plot showcasing dominance of different period frequencies in cycles per year. By examining the height and location of the peaks in the periodogram, we can determine which cycles are most prominent and obtain a rough estimate of their lengths. This serves as a baseline before applying smoother or more advanced techniques.

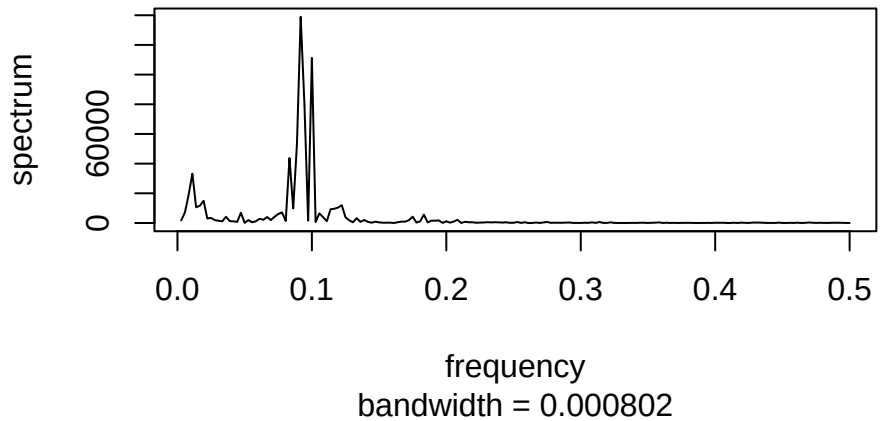


Figure 4: Raw Periodogram for Sunspot Data

In the figure above, we can see a spike that is most dominant at and slightly before the frequency of 0.1 cycles per year. The plot is, however, very raw and it is unclear as to which frequency is the underlying one. To obtain a better visualization, we will use simple smoothing for the periodogram.

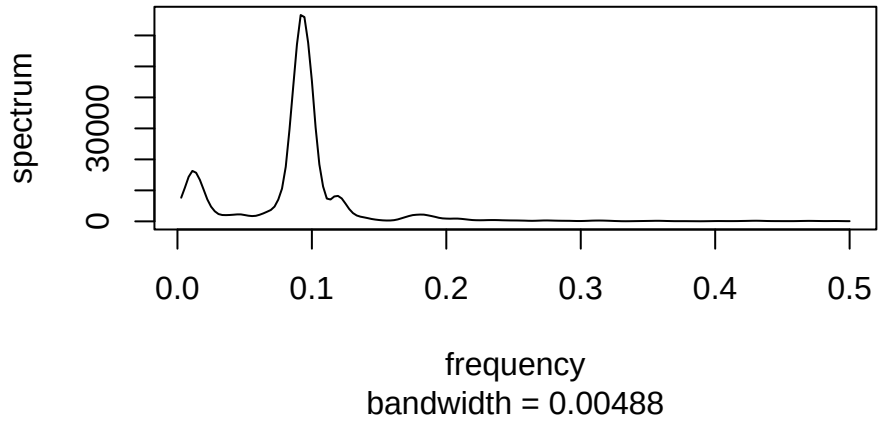


Figure 5: Smoothed Periodogram

With the smoothed periodogram, the dominant peak becomes much easier to see. The main frequency is slightly below 0.1 cycles per year, which corresponds to a period of roughly 11 years, consistent with what is known about sunspot activity.

Now, using the frequency data, we estimated dominant frequency of the series by identifying the value of f at which the spectral density $S(f)$ reaches its maximum:

$$f^* = \arg\max_f S(f).$$

The corresponding cycle length was then obtained by inverting this frequency:

$$T = \frac{1}{f^*} = 10.90909 \ years$$

(The code used for this calculation can be found in the Analysis C section of the Appendix)

AR Spectral Density

For appropriate procedure, we also wanted to examine the spectral density of this data by fitting an AR(p) model to the time series. It offers a parametric perspective, allowing us to assess whether the dominant periodicity remains consistent under a fitted time-series model.

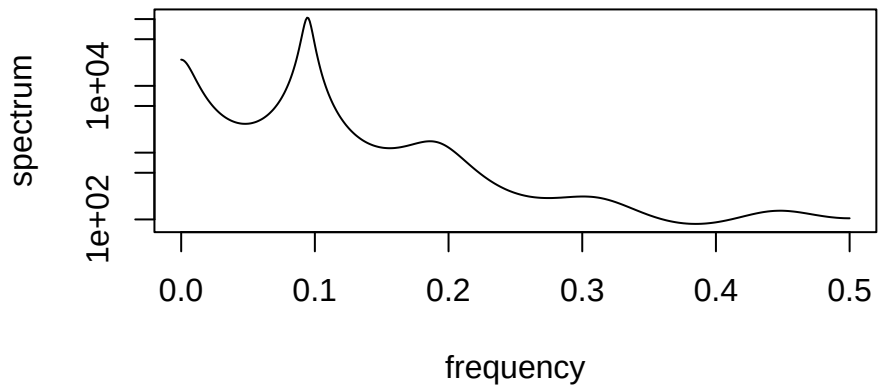


Figure 6: AR Spectral Density Estimate

The plot above also clearly shows a peak at slightly below 0.1 cycles per year, leading to the conclusion of the underlying period being approximately 11 years. However, while the spectral density plot does perform better than the raw periodogram, it essential leads to the same result as a smoothed periodogram while also requiring stationarity, making it useful only in specific situations, which this study is an example of. Thus, this approach serves only as a helpful confirmation rather than a different insight into the data.

Complex Demodulation

Since based on the time series plot the data looked like a drifting oscillation, and the referenced paper also mentioned using it, we wanted to try implementing complex demodulation as one of the methods to extract the instantaneous period. Berger A., et al, however, did not use it to approximate the period and only used it for amplitude and phase estimation. They also mentioned that the 11-year-quasi-peridoicity was "highly unstable" (Berger et al., 138).

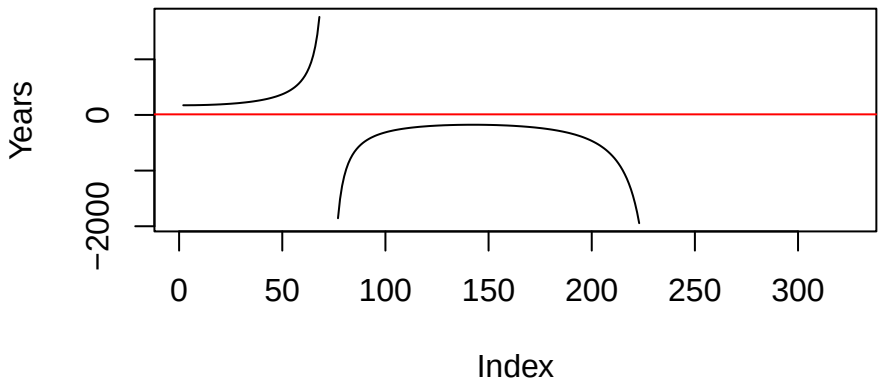


Figure 7: Instantaneous Period of Sunspot Cycle (11 Years)

As can be seen in the plot above, both curves diverge to infinity and negative infinity, suggesting that the 11 year data is, in fact, very unstable, even with multiple attempts at smoothing the phase. Now, since the paper mentioned the 22-year period being more stable, below are the findings as well.

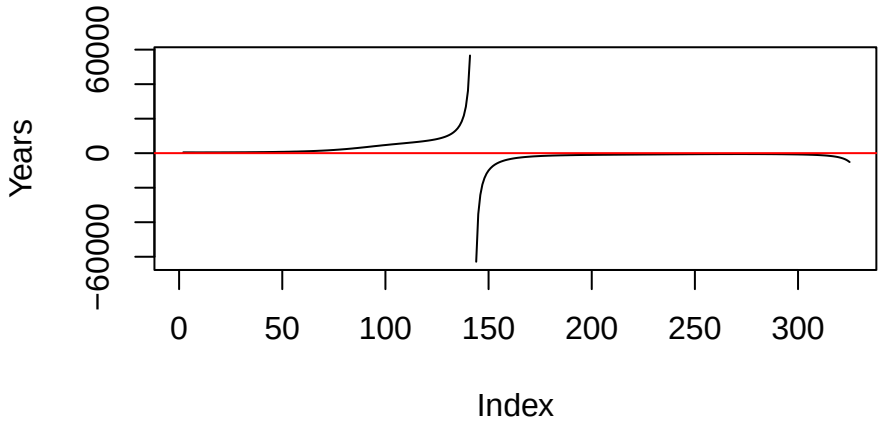


Figure 8: Instantaneous Period of Sunspot Cycle (22 Years)

Now, considering the 22-year period as the period of interest, the instantaneous period still appears quite unstable with both curves still diverging to inifinity. If the data was a smooth oscillation with gradually varying period, we certainly could have found a clear estimate of both period and the change in amplitude over time, since across centuries there would definitely have been some variance. However, the series is anharmonic, nonlinear, and contains multiple interacting cycle, meaning an instantaneous frequency is not exactly defined, so complex demodulation shows unstable results regardless of the defined period being 11 or 22 years.

Wavelet

Traditional spectral methods, like the periodogram and AR-based estimates, assume that the underlying period is roughly constant over time. However, the sunspot series clearly shows changes in both amplitude and cycle length across the years, suggesting that a single global spectrum may not fully describe its behavior. To capture how the dominant periodicity changes over time, we will use wavelet analysis.

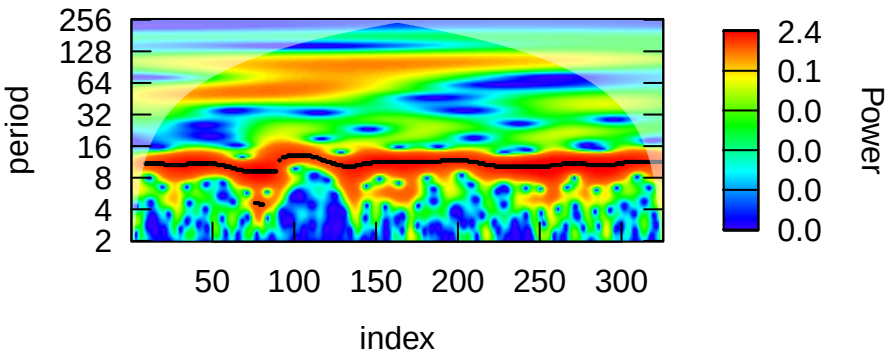


Figure 9: Wavelet-Based Evolutive Harmonic Analysis

The wavelet power spectrum shows a clear concentration of energy around periods close to 10–12 years across the entire series, confirming the well-known sunspot cycle. The thickness of the red band suggests that the dominant cycle length is not perfectly constant but changes over time. Higher-period components (around 30–120 years, seen with the orange band) also appear with weaker but noticeable power. Overall, the wavelet plot highlights both the persistent approximate 11-year cycle and its gradual fluctuations across centuries.

For each time index t, let $P(\tau, t)$ be the wavelet power at period τ.
The dominant period at time t is defined as

$$\tau^*(t) = \arg\max_{\tau} P(\tau, t).$$

The overall average dominant period is then computed as

$$\overline{\tau} = \frac{1}{T}\sum_{t=1}^{T} \tau^*(t) = 10.9 \ \ years,$$

where T is the number of time points with a well-defined dominant period. (The code used for this calculation can be found in the Analysis C section of the Appendix)

Multitaper

In this project, we were also interested in how the dominant cycle changes over time rather than just identifying a single global period. While the methods we used earlier provide useful information, they do not capture local variations in the cycle length very well. The multitaper approach offers a more stable way to estimate the spectrum within moving windows, allowing us to track fluctuations in the sunspot period across centuries. Using this method, we can visualize how the estimated cycle length evolves over time, as shown in the following plot.

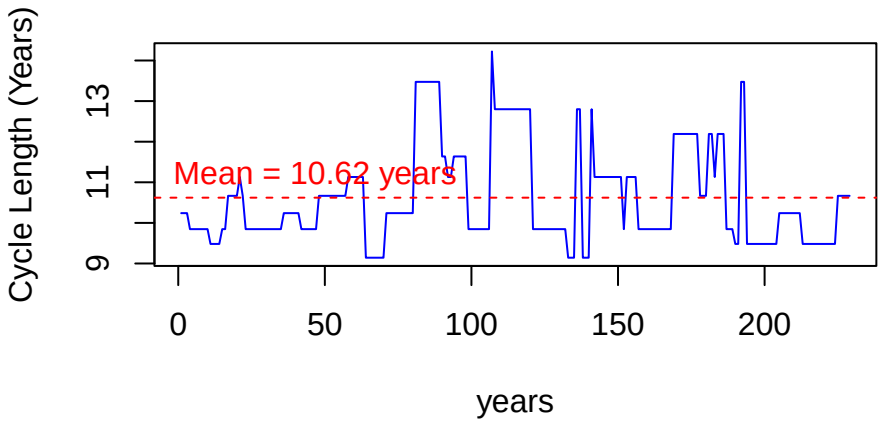


Figure 10: Multitaper Time-Varying Period Estimate

The multitaper rolling estimate shows that the dominant sunspot cycle varies over time, generally staying within the 9–13 year range. While the cycle is roughly centered near the expected value, noticeable fluctuations occur across different windows. The dashed red line marks the overall average cycle length of approximately 10.62 years, highlighting that the cycle is stable and consistent with the known information on average but not perfectly constant.

Discussion & Conclusion

Outside of complex demodulation, which was not quite suited for period estimation in this particular study, all models strongly agree on the dominant cycle length of approximately 11 years, showing that the period has not shifted and is consistent with past research. While this is well-known information, understanding how it is obtained was still important both for astronomy, where accurate cycle estimates inform of solar activity, and statistics, where this dataset serves as a great example of periodic behaviour.

When looked at on their own, the most useful approaches were Wavelet Analysis and Multitaper Analysis, as they both provided the dominant period and clearly showed variations of said period across the windows of time and how strong those variations were.

Overall, however, while all of the tools did show us something useful, all of them complement one another. Together, they provide a more complete understanding for the underlying period of the sunspot solar cycle

and show why multiple spectral tools are usually needed to fully characterize complex and, in this case, astronomical time series.

Now, there is not much to be done in terms of improvements for future research, since this topic is already quite well-studied. In theory, more data would allow a sharper estimate of the solar cycle's true period, but collecting several additional centuries of observations may not be practical within the typical research timeline.

References

Berger A., et al. Evolutive Spectral Analysis of Sunspot Data over the past 300 years. Royal Society, 1990. https://www.jstor.org/stable/53602.

Dangi, S. (2023). CO2 emissions across countries, regions, and sectors [Data set]. Kaggle. https://www.kaggle.com/datasets/shreyanshdangi/co-emissions-across-countries-regions-and-sectors

National Centers for Environmental Information. (2023). National climate report: December 2023. NOAA. https://www.ncei.noaa.gov/news/national-climate-202312

National Centers for Environmental Information. (2024). National climate report: December 2024. NOAA.https://www.ncei.noaa.gov/news/national-climate-202413

Nguyen, G. B. (2023). Average temperature from 1900 to 2023 [Data set]. Kaggle. https://www.kaggle.com/datasets/giabchnguyn/average-temperature-from-1900-to-2023

Ritchie, H., Roser, M., & Rosado, P. (2024). CO2 and greenhouse gas emissions. Our World in Data. https://ourworldindata.org/co2-emissions

Royal Observatory of Belgium. Wolf Sunspot Number Data. SIDC. https://www.sidc.be/SILSO/datafiles

United Nations. (n.d.). Climate change. United Nations. https://www.un.org/en/global-issues/climate-change

Data Dictionary Analysis A

Table 4: Data Dictionary for Variables Used in Analysis A

Variable Name	Units	Missing Values	Source
year	Year (annual)	None	Our World In Data
population	Number of people	None	Our World In Data
gdp	Inflation-and cost-of-living-adjusted USD	None	Our World In Data
co2 emissions	Million metric tons	None	Our World In Data
methane emissions	Metric tons	None	Our World In Data
nitrous oxide emissions	Metric tons	None	Our World In Data
average temperature (F)	Degrees Fahrenheit (annual average)	None	NCEI

Data Dictionary Analysis C

Table 5: Data Dictionary for Variables Used in Analysis C

Variable Name	Units	Missing Values	Source
Year	Year (annual)	None	Royal Observatory of Belgium
Sunspots	Mean annual sunspot count	None	Royal Observatory of Belgium

Appendix Analysis A

Levene's Test

```
# Rolling variance plot (window = 100 years)
roll_var <- rollapply(df$Average_Fahrenheit_Temperature, width=10, FUN=var, by=1, align="right")

mid <- floor(nrow(df)/2)
var_first_half <- var(df$Average_Fahrenheit_Temperature[1:mid])
var_second_half <- var(df$Average_Fahrenheit_Temperature[(mid+1):nrow(df)])
var_first_half
```

```
## [1] 1.99624
```

```
var_second_half
```

```
## [1] 2.064728
```

```
# Levene's Test
df$period <- ifelse(df$year <= median(df$year), "Early", "Late")
leveneTest(Average_Fahrenheit_Temperature ~ period, data=df)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.2198   0.64
##       121
```

Regression Models Summaries

Full Model

```
# Full model: time + population + gdp + co2 + nitrous oxide + methane
model_full <- lm(Average_Fahrenheit_Temperature ~ year + population + gdp + co2 + nitrous_oxide + methan
summary(model_full)
```

```
##
## Call:
## lm(formula = Average_Fahrenheit_Temperature ~ year + population +
##     gdp + co2 + nitrous_oxide + methane, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1745 -0.7946  0.0780  0.7971  3.9336
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.242e+02  1.123e+02   1.106   0.2709
## year          -3.835e-02  6.065e-02  -0.632   0.5284
```

```
## population     1.191e-08  4.906e-08   0.243   0.8087
## gdp            6.077e-14  3.310e-13   0.184   0.8546
## co2            8.450e-04  4.256e-04   1.985   0.0495 *
## nitrous_oxide  5.365e-03  1.026e-02   0.523   0.6021
## methane       -5.573e-03  5.243e-03  -1.063   0.2900
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.258 on 116 degrees of freedom
## Multiple R-squared:  0.3603, Adjusted R-squared:  0.3272
## F-statistic: 10.89 on 6 and 116 DF,  p-value: 1.36e-09
```

Social Model

```
# Socioeconomic only: time + population + gdp
model_socio <- lm(Average_Fahrenheit_Temperature ~ year + population + gdp, data=df)
summary(model_socio)
```

```
##
## Call:
## lm(formula = Average_Fahrenheit_Temperature ~ year + population +
##     gdp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3882 -0.7986  0.0693  0.7397  3.8433
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.140e+02  7.747e+01   1.471   0.1439
## year        -3.299e-02  4.160e-02  -0.793   0.4293
## population   8.481e-09  2.494e-08   0.340   0.7344
## gdp          2.260e-13  1.190e-13   1.900   0.0599 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.269 on 119 degrees of freedom
## Multiple R-squared:  0.3315, Adjusted R-squared:  0.3147
## F-statistic: 19.67 on 3 and 119 DF,  p-value: 2.002e-10
```

Emissions Model

```
# Emissions only: time + co2 + nitrous oxide + methane
model_emissions <- lm(Average_Fahrenheit_Temperature ~ year + co2 + nitrous_oxide + methane, data=df)
summary(model_emissions)
```

```
##
## Call:
## lm(formula = Average_Fahrenheit_Temperature ~ year + co2 + nitrous_oxide +
##     methane, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2973 -0.8016  0.0035  0.7114  3.8054
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   64.1905382 22.0135478   2.916 0.004245 **
## year          -0.0061699  0.0115945  -0.532 0.595631
## co2            0.0010566  0.0002966   3.563 0.000531 ***
## nitrous_oxide  0.0123107  0.0078685   1.565 0.120365
## methane       -0.0096345  0.0026997  -3.569 0.000520 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.254 on 118 degrees of freedom
## Multiple R-squared:  0.3533, Adjusted R-squared:  0.3314
## F-statistic: 16.12 on 4 and 118 DF,  p-value: 1.477e-10
```

Model Comparisons

```
# Compare models
AIC(model_full, model_socio, model_emissions)
```

```
##                 df      AIC
## model_full       8 414.2226
## model_socio      5 413.6341
## model_emissions  6 411.5568
```

```
BIC(model_full, model_socio, model_emissions)
```

```
##                 df      BIC
## model_full       8 436.7201
## model_socio      5 427.6950
## model_emissions  6 428.4299
```

```
paste0(summary(model_full)$adj.r.squared, " model full adj r squared")
```

```
## [1] "0.327229535241707 model full adj r squared"
```

```
paste0(summary(model_socio)$adj.r.squared, " model socio adj r squared")
```

```
## [1] "0.314693543392864 model socio adj r squared"
```

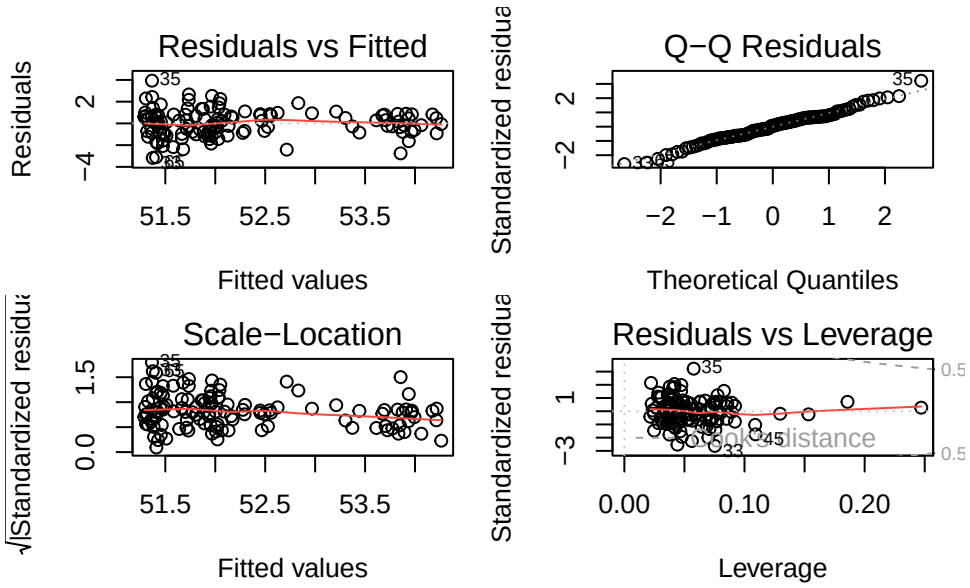
```
paste0(summary(model_emissions)$adj.r.squared, " model emissions adj r squared")
```

```
## [1] "0.331419716969196 model emissions adj r squared"
```

Regression and Residual Analysis

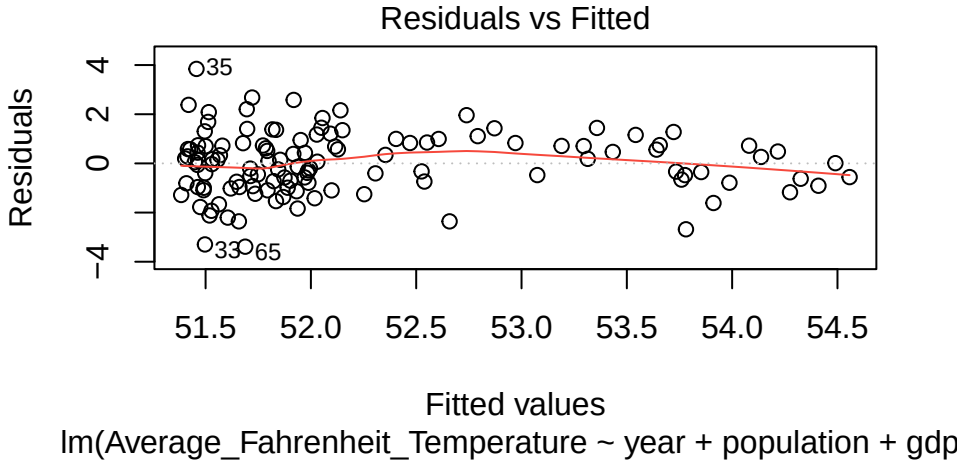
Full Model

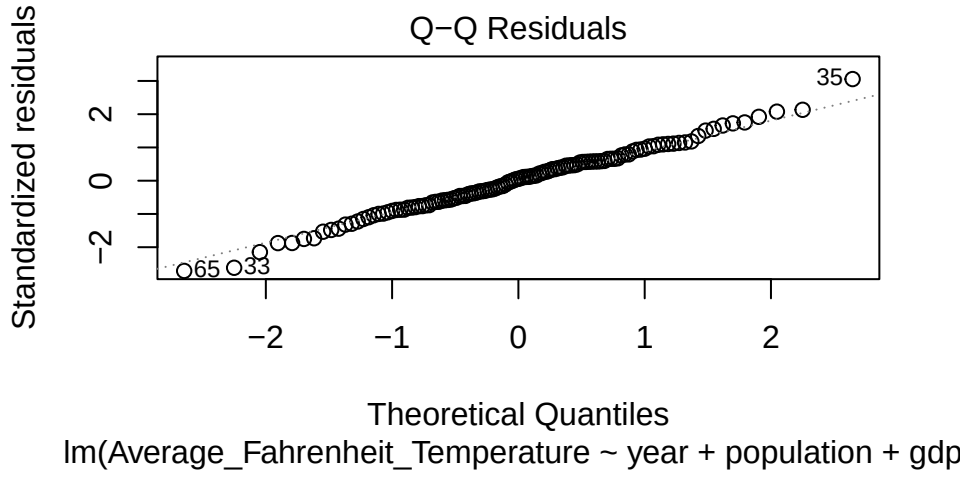
```
# Residual plots
par(mar = c(4,4,2,1))
par(mfrow=c(2,2))
plot(model_full)
```

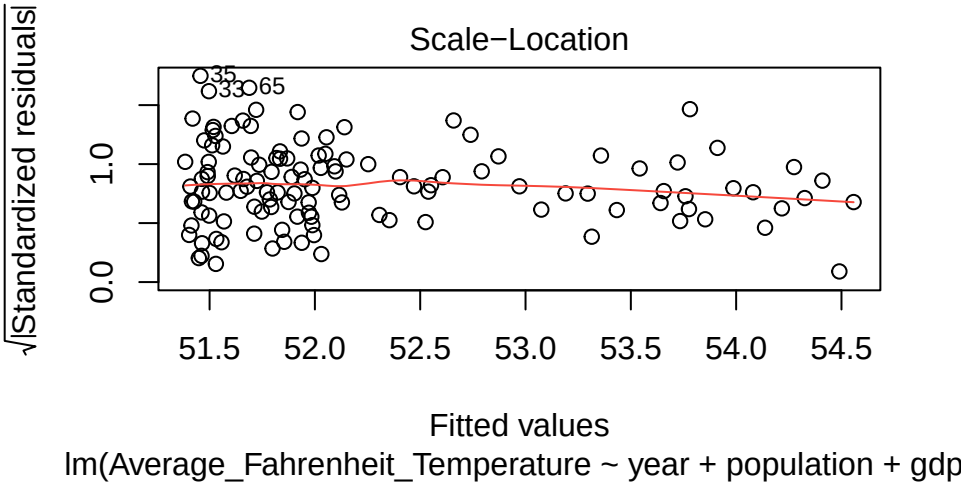


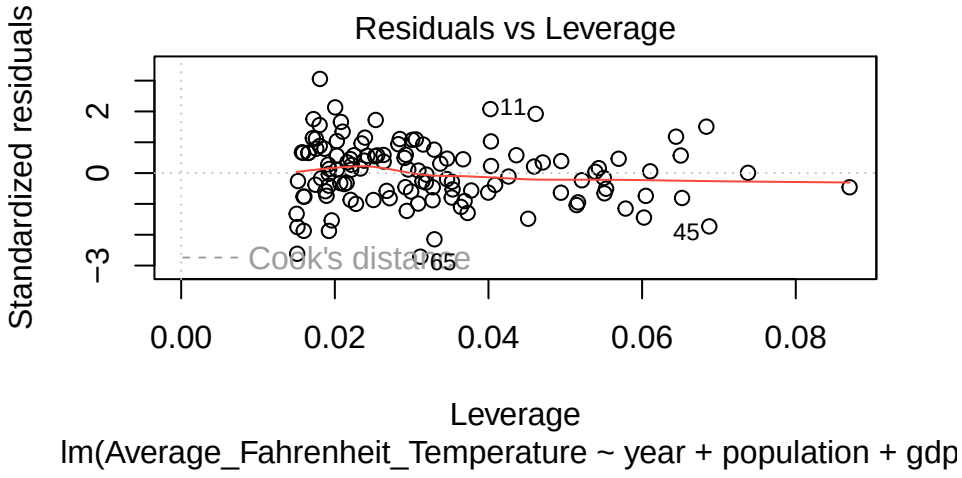
Social Model

```
plot(model_socio)
```



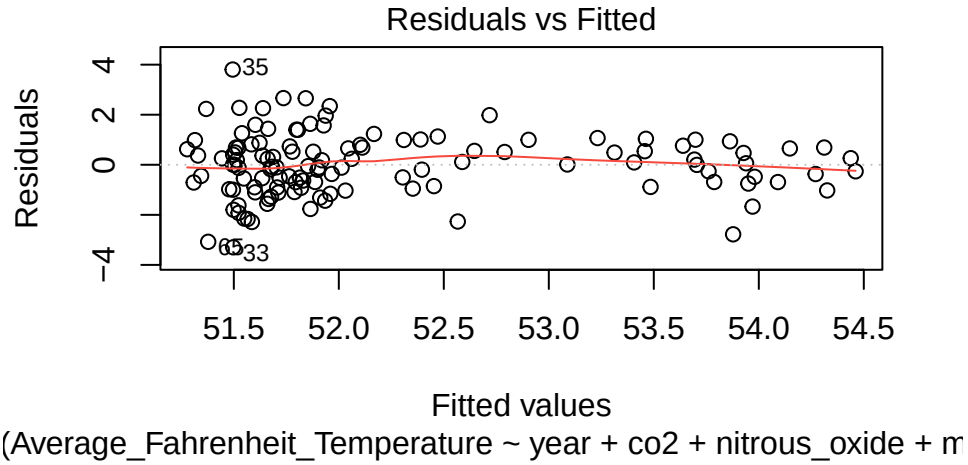

Q−Q Residuals
Standardized residuals
35
65
33
−2
−1
0
1
2
Theoretical Quantiles
lm(Average_Fahrenheit_Temperature ~ year + population + gdp

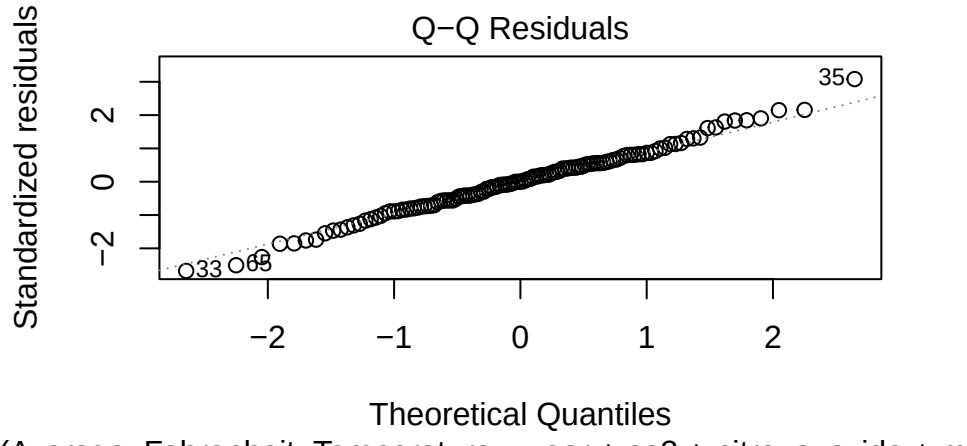

Scale−Location
√|Standardized residuals|
35
33
65
0.0
1.0
51.5
52.0
52.5
53.0
53.5
54.0
54.5
Fitted values
lm(Average_Fahrenheit_Temperature ~ year + population + gdp

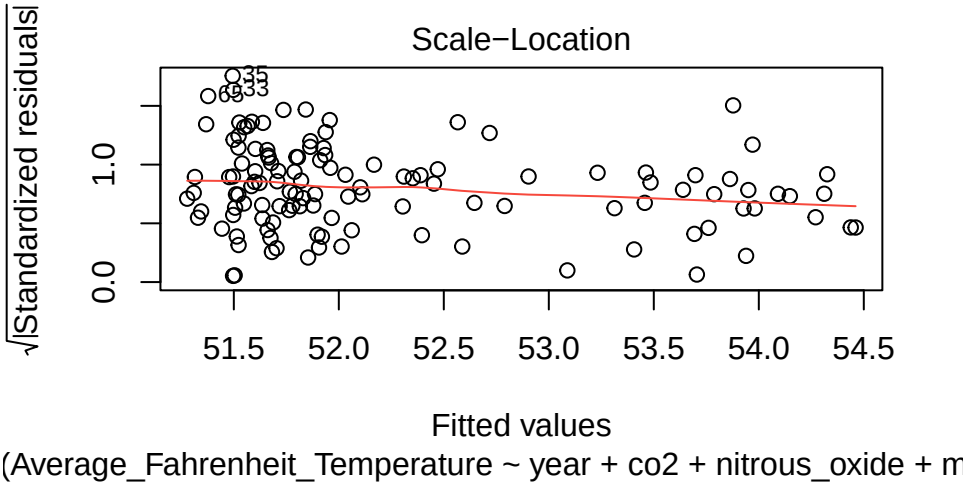


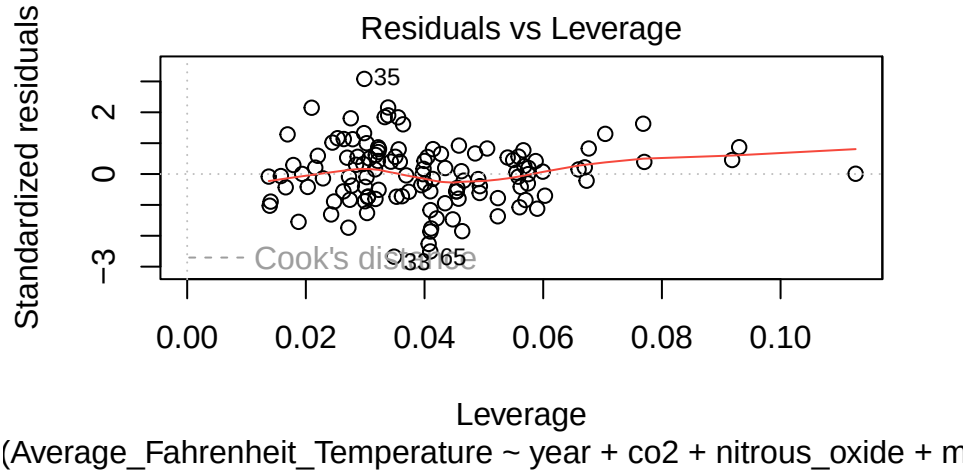
Emissions Model

```
plot(model_emissions)
```




Q−Q Residuals
Standardized residuals
35
33
69
−2
0
2
−2
−1
0
1
2
Theoretical Quantiles
(Average_Fahrenheit_Temperature ~ year + co2 + nitrous_oxide + m

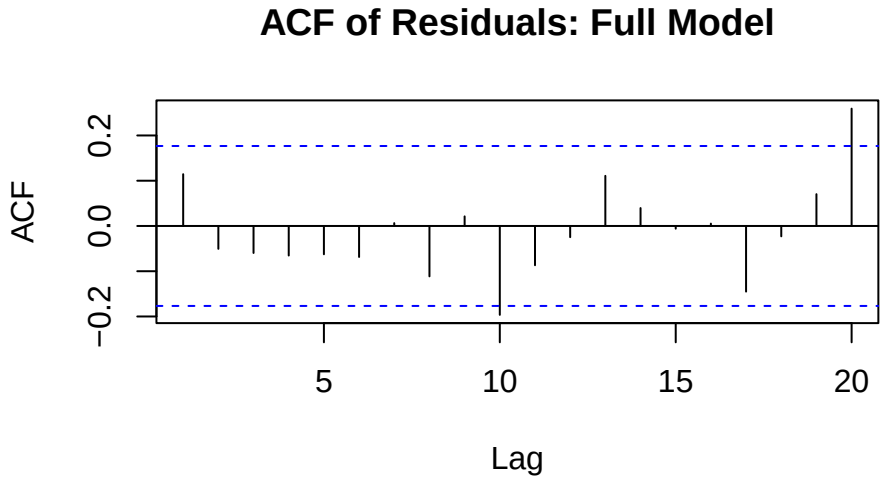

Scale−Location
√|Standardized residuals|
35
69
33
0.0
1.0
51.5
52.0
52.5
53.0
53.5
54.0
54.5
Fitted values
(Average_Fahrenheit_Temperature ~ year + co2 + nitrous_oxide + m



Residual Analysis and ACF plots

Full model

```
# Residual autocorrelation
acf(residuals(model_full), main="ACF of Residuals: Full Model")
```

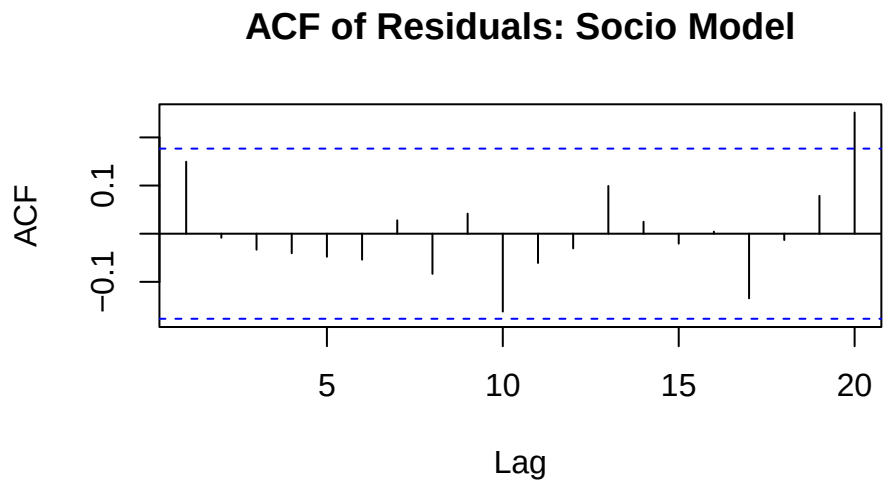



```
Box.test(residuals(model_full), type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  residuals(model_full)
## X-squared = 1.65, df = 1, p-value = 0.199
```

Social Model

```
acf(residuals(model_socio), main="ACF of Residuals: Socio Model")
```

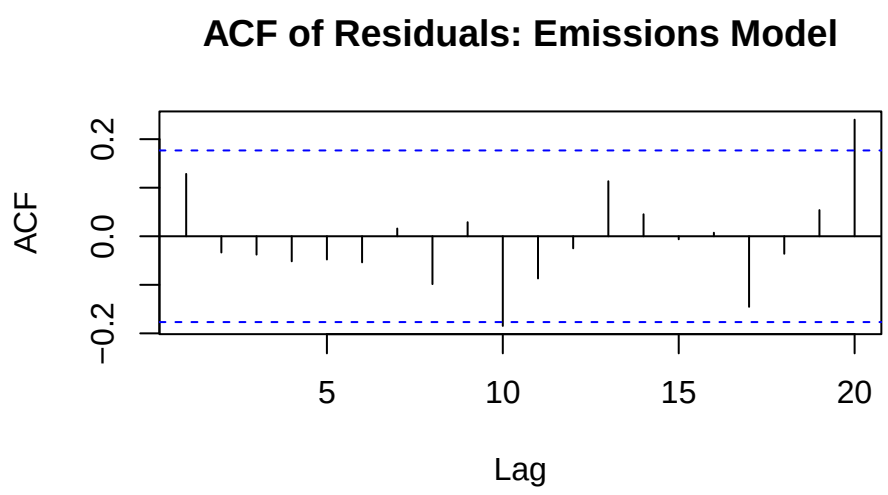



```
Box.test(residuals(model_socio), type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  residuals(model_socio)
## X-squared = 2.8156, df = 1, p-value = 0.09335
```

Emissions Model

```
acf(residuals(model_emissions), main="ACF of Residuals: Emissions Model")
```




```
Box.test(residuals(model_emissions), type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  residuals(model_emissions)
## X-squared = 2.0776, df = 1, p-value = 0.1495
```

Forecasting future 5 values

```
summary(df$population)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##  74829905 120230362 183489491 191465776 258038722 341534041
```

```
summary(df$gdp)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 6.140e+11 1.310e+12 3.339e+12 6.004e+12 9.402e+12 1.949e+13
```

```
summary(df$co2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   662.7  1741.6  2897.3  3389.8  5070.9  6132.2
```

```
summary(df$nitrous_oxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   63.23   78.91  200.50  180.12  259.54  303.57
```

```r
summary(df$methane)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   208.5   362.0   653.4   585.5   794.5   877.2
```

```r
last_gdp <- tail(na.omit(df$gdp), 1)

future_years <- data.frame(
  year = (max(df$year) + 1):(max(df$year) + 5),
  co2 = tail(df$co2,1) * (1 + 0.01)^(1:5),
  nitrous_oxide = tail(df$nitrous_oxide,1) * (1 + 0.01)^(1:5),
  methane = tail(df$methane,1) * (1 + 0.01)^(1:5)
)

str(future_years)
```

```
## 'data.frame':    5 obs. of  4 variables:
##  $ year         : int  2023 2024 2025 2026 2027
##  $ co2          : num  5130 5181 5233 5285 5338
##  $ nitrous_oxide: num  252 255 257 260 262
##  $ methane      : num  702 709 716 724 731
```

```r
any(is.na(future_years))
```

```
## [1] FALSE
```

```r
final_model <- model_emissions

future_forecast <- predict(final_model, newdata = future_years, interval = "prediction")
future_forecast
```

```
##        fit      lwr      upr
## 1 53.46812 50.87401 56.06224
## 2 53.47955 50.88533 56.07377
## 3 53.49115 50.89684 56.08545
## 4 53.50292 50.90856 56.09728
## 5 53.51488 50.92049 56.10926
```

Appendix Analysis C

Sunspot Time Series ADF Test

```r
adf.test(data_ts, )
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_ts
## Dickey-Fuller = -5.3806, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

Smoothed Periodogram Period

```
spec <- spec.pgram(x, plot=FALSE)
freq_max <- spec$freq[which.max(spec$spec)]
period_est <- 1 / freq_max
```

Wavelet Period

```
power <- result$Power
periods <- result$Period
dominant_periods <- apply(power, 2, function(col) {
  periods[which.max(col)]
})

mean_dominant_period <- mean(dominant_periods, na.rm = TRUE)
```