Table of Contents

1. Introduction

1.1 Topic

Throughout the course of history, Earth's climate has been subject to different changes. Over the past 2 million years, there have been sequences of glacial periods, the last one having occurred about 11,700 years ago. This marked the beginning of the modern climate era (Clayton et al., 2006).

The reason for these climate changes was studied by many different scholars, though the most accredited theory is the one proposed a century ago by a Serbian scientist, Milutin Milankovitch. He hypothesized that there are regular changes in Earth's orbit and the relative position to the Sun that alters the solar radiation of our planet (Buis, 2021). These cycles are key drivers of Earth's long-term climate and are considered to be the main causes of the glaciation periods, but they cannot explain the current period of rapid warming that our planet is facing. Today, it is seen how "unequivocal human influence has warmed the atmosphere, ocean and land. Widespread and rapid changes in the atmosphere, ocean, cryosphere and biosphere have occurred" (The Intergovernmental Panel on Climate Change [IPCC], 2021).

1.2 Purpose and Research Question

The purpose of the project is to both to provide evidence that the average global temperature is increasing and to analyze factors having a high impact on climate temperature. In this paper, a probabilistic model is developed to describe global warming phenomena, rather than build a model capable of learning from mistakes, i.e. model with prediction purposes.

The study will explore, explain and visualize various data sources and datasets to assess what is the evidence, causes, and effects of global climate change. Moreover, using different statistical methods the study will evaluate the main cause of climate change. The findings will be narrowed to understand how different drivers affect climate change, taking into consideration both human and natural factors.

Overall, the project aims to give an answer to the following research questions:

- Is there evidence that global yearly average temperatures are increasing?
- Which are the main drivers influencing climate change ?

2. Data Sources

Data used for the ongoing of the project have been retrieved from three websites: NASA's Goddard Institute for Space Studies, Berkeley Earth and Eurostat

2.1 NASA's Goddard Institute for Space Studies

The first source is The NASA Goddard Institute for Space Studies (GISS) for global land-ocean temperature. The institute uses analysis of datasets with global models of atmospheric, land, surface and oceanic processes along with analysis of past climate change on Earth and other planetary atmospheres in order to predict atmospheric and climate changes.

The GISS surface temperature analysis is an estimate of global surface temperature change that began in the late 1970's. At the time, stations were grouped into 80 equal area boxes, the various anomaly series in a box were combined into a single anomaly series; these then were averaged across each of eight latitude belts. The global mean was estimated from an area weighing of the latitudinal means. The analysis will examine annual global land-ocean temperature anomalies.

2.2 Berkeley Earth

The second data source, Berkeley Earth, provided the monthly anomaly temperature by continent. It is an independent and non-governmental American organization specialized in climate science. The firm provides data for the purpose to serve a wide range of audiences from politicians to students. Accuracy and objectivity in their data make it widely used as information or as input for research analysis.

Their method took temperature observations from a large collection of weather monitoring stations and produced an estimate of the underlying global temperature field across all of the Earth's land areas. This analysis will examine monthly land-surface temperature averages (in Celsius), expressed as anomalies relative to the January 1951-December 1980 average.

The Berkeley Earth study ran every month from 1750 until present times. They provide monthly data for different continents. For instance, taking into consideration Europe, Berkeley Earth uses approximately 5790 temperature stations, resulting in around 2.8 million monthly observations. Moreover, the base estimated temperature in Europe (Jan 1951-Dec 1980 average temperature) is 8.15 °C +/- 0.12 °C.

When the station coverage within the region became too low, Berkeley Earth research reported the values as missing (i.e NaN). However, when at least 75% of the necessary values over a specific interval are available, time averages will be reported. The Berkley Earth framework is expected to be robust against most forms of bias; however, the impact of some forms of possible systematic bias is still being studied by the institute.

2.3 Eurostat

Eurostat has been added as a source to this project after completing the first two assignments. The aim of the study was to build probabilistic models designed to describe the cause of global warming phenomena, and Eurostat provides different data related to the causes of climate change that are worth investigating.

Eurostat is a directorate-General of the European Commission and is in charge of providing accurate data and statistical information at European level to various EU institutions. The website contains a vast array of datasets for a number of drivers. In the analysis the possible drivers to global warming considered are related to greenhouse gas emissions, energy, transport, industrial processes, and waste. After analyzing all the datasets relative to those topics, only a few of them were considered to run the analysis. The logic behind this approach was on one hand to avoid using datasets with much missing data, and on the other hand only select datasets with enough input, i.e. data collected over several years. Sorting those datasets permitted us to select the drivers regarding greenhouse gas emissions, energy, transport, industrial processes, and waste, and then to proceed with creating variables which in the end will be put into models.

3. Dataset Description & Exploration

3.1 Descriptive statistics

NASA/GISS data

The following NASA data are plain-text files in tabular format of temperature anomalies, i.e. deviations from the corresponding 1951-1980 means.

The NASA/GISS dataset has 141 rows and 3 columns. Each row indicates the year, the annual anomaly value, and the Lowess smoother value. The output of the *str()* function highlights that the "*Year*" is an integer variable while "*No_Smoothing*", and "*Lowess.5*" are numeric variables.

Year	No_Smoothing	Lowess.5.
Min. :1880	Min. :-0.48000	Min. :-0.41000
1st Qu.:1915	1st Qu.:-0.20000	1st Qu.:-0.22000
Median :1950	Median :-0.07000	Median :-0.04000
Mean :1950	Mean : 0.04858	Mean : 0.04858
3rd Qu.:1985	3rd Qu.: 0.23000	3rd Qu.: 0.22000
Max. :2020	Max. : 1.02000	Max. : 1.01000

Figure 1. Descriptive statistics for NASA/GISS dataset

In this table, it is possible to see that the global time-series data are available from 1880 to 2020. The mean of the global yearly anomaly temperature is 0.04 °C with a maximum value of 1.02 °C and a minimum one of -0.48 °C. Lowess values lead to a fitting line to the time series data plot with noisy data values and sparse data points.

Berkeley Earth data

Temperatures are in Celsius and reported as anomalies relative to the Jan 1951-Dec 1980 average. Uncertainties represent the 95% confidence interval for statistical and spatial undersampling effects. As Earth's land is not distributed symmetrically about the equator, there exists a mean seasonality to the global land-average. For each month, the estimated land-surface average is reported for that month and its uncertainty. The corresponding values for year, five-year, ten-year, and twenty-year moving averages centered about that month (rounding down if the center is in between months) are also reported.

The resulting dataset has 1158 rows and 4 columns. Each row indicates the annual anomaly value, the Lowess smoother value, and the relative zone. The output of the *str()* function highlights that the "*zone*" is a character variable while "*year*", "*anomaly*" and "*lowess*" are numeric variables.

year	yearly_anomaly	lowess	zone
Min. :1750	Min. :-1.684333	Min. :-0.769483	Length:1158
1st Qu.:1887	1st Qu.:-0.402314	1st Qu.:-0.369790	Class :character
Median :1935	Median :-0.096292	Median :-0.120499	Mode :character
Mean :1928	Mean :-0.009384	Mean :-0.006192	NA
3rd Qu.:1979	3rd Qu.: 0.337562	3rd Qu.: 0.176794	NA
Max. :2020	Max. : 2.163750	Max. : 1.750654	NA

Figure 2. Descriptive statistics for Berkeley Earth data

In this table, it is possible to see that the time-series data are available from 1750 to 2020. The mean of the yearly anomaly temperature is -0.009 °C with a maximum value of 2.163 °C and a minimum one of -1.684 °C. Lowess values lead to a fitting line to the time series data plot with noisy data values and sparse data points. Zone is the only variable expressed in a character format, it expresses the continent of the observation.

3.1 Variables description

Eurostat data

The Eurostat database was used to gather data concerning the independent variables related to potential drivers influencing climate change. The data selected for the model belongs to different sectors, namely: Greenhouse Gas Emissions, Energy, Transport, Industrial Processes and Product Use and Waste (Figure 3).

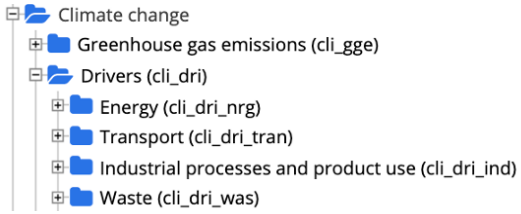

Figure 3. Folder contains data regarding climate change (Eurostat)

Source. Eurostat database

Each folder of the different sectors of climate change contained several datasets concerning various variables of that specific sector. Taking as an example Greenhouse Gas Emissions it is possible to notice that not all the dataset were considered for the final model, more specifically only "Greenhouse gas emissions by source sector", "Greenhouse gas emissions from agriculture", "Greenhouse gas emissions intensity of energy consumption" and "Average CO2 emissions per km from new passenger cars" were considered in the statistical analysis (Figure 4).

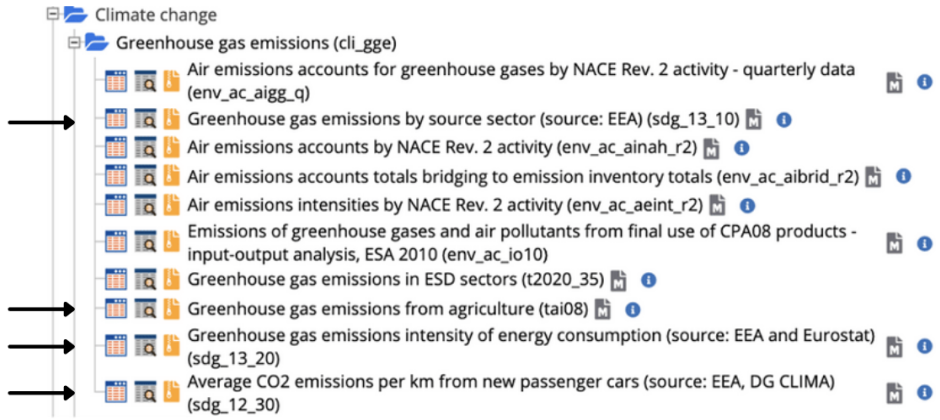

Figure 4. Dataset related to greenhouse gas emission (Eurostat)

Source. Eurostat database

The factors analyzed to choose the datasets were: (1) the sample size, (2) the amount of missing data and (3) the affinity with the research question. In the Appendix D it is possible to see all the datasets for every sector examined comprehensively of both the rejected and the accepted ones, as well as a detailed description of each database. In Appendix B it is possible to see the R code used to gather, transform and merge all the data sets considered.

The following variables were in the selected datasets and are used in the model. This gives further understanding and context to the decision making process of how they were chosen.

Greenhouse gas emissions

This indicator tracks trends in greenhouse gas (GHG) emissions by agriculture, estimated and reported under the United Nations Framework Convention on Climate Change (UNFCCC), the Kyoto Protocol and EU Regulation (Decision 525/2013/EC).

- *tai08_CO2_PC_CRF3_EU28*: Percentage of Carbon dioxide gas emission from agriculture.

Waste

Those variables are related to municipal waste by waste management operations. It is the waste from households, offices, commerce and public institutions and it excludes waste from agriculture and industries. The waste is collected by municipal authorities and then undergoes various treatment such as incineration, recycling, composting and landfilling. Municipal waste is classified by waste generation and by the treatment operations.

- *env_wasmun_GEN_KG_HAB_EU28*: Represents the total waste generation in kilograms per capita;

- *env_wasmun_TRT_KG_HAB_EU28*: Represents the total waste treatment in kilograms per capita.

Industrial Processes and Product Use

Those variables are expressed as indices and are related to production in various industries. Construction indices are broken down by Classification of Types of Constructions (CC).

- *sts_copr_a_PROD_F_CC1_CA_I10_EU28*: Volume index relative to 2010 (=100) in production in constructions of buildings;
- *sts_copr_a_PROD_F_CC2_CA_I10_EU28*: Volume index relative to 2010 (=100) in production in constructions in civil engineering works;
- *sts_inpr_a_PROD_C_CA_I10_EU28*: Volume index relative to 2010 (=100) in production of industry relative to manufacturing;
- *sts_inpr_a_PROD_D_CA_I10_EU28:* Volume index relative to 2010 (=100) in production in Industry relative to electricity, gas, steam and air conditioning supply.

Transport

Modal split of passenger in transport is defined as the percentage of transport by passenger cars, buses and coaches, and trains in total inland passenger transport performance, measured in passenger-kilometers:

- *tran_hv_psmod_PC_BUS_TOT_EU28*: Relative to motor coaches, buses, and trolley buses;
- *tran_hv_psmod_PC_TRN_BUS_TOT_AVD_EU28*: Relative to trains, motor coaches, buses and trolley buses.

Volume of passenger transport relative to GDP is expressed as the ratio between the total transport performance of passengers using the inland modes (road and rail):

- *tran_hv_pstra_I10_EU28:* Volume index relative to 2010 (=100) of passenger transport relative to GDP;
- *ttr00005_TOT_LOADED_THS_T_EU28*: Thousands of tons of carriage goods loaded by road transport.

Energy

The indicator measures the total energy needs of a country excluding all non-energy use of energy carriers (e.g. natural gas used not for combustion but for producing chemicals).

- *t2020_rk200_TOTAL_KTOE_FC_OTH_HH_E_EU28*: Final energy consumption in households in thousand tonnes of oil equivalent;

Final energy consumption in thousand tonnes of oil equivalent by:

- *ten00123_FC_E_C0000X0350.0370_KTOE_EU28*: Solid fossils fuels;
- *ten00123_FC_E_C0350.0370_KTOE_EU28*: Manufactured gses;
- *ten00123_FC_E_E7000_KTOE_EU28:* Electricity;
- *ten00123_FC_E_H8000_KTOE_EU28:* Heat;
- *ten00123_FC_E_04000XBIO_KTOE_EU28:* Oil and petroleum product (excluding biofuel portion);
- *ten00123_FC_E_S2000_KTOE_EU28:* Oil shale and oil sands;
- *ten00123_FC_E_TOTAL_KTOE_EU28* in total;
- *ten00123_FC_E_W6100_6220_KTOE_EU28:* Non-renewable waste.

3.2 Visualization

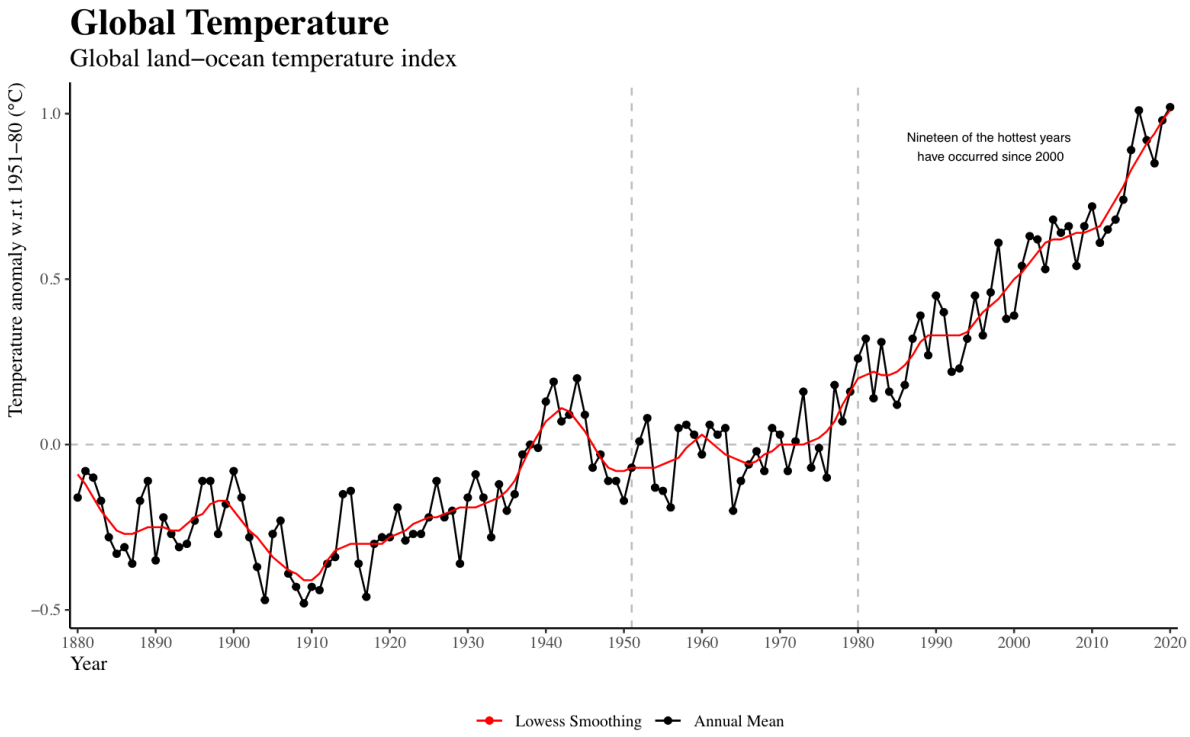

Figure 5. Global temperature anomaly

Figure 5 illustrates the change in global land-ocean temperature relative to 1951-1980 average temperature. See appendix A for R code.

Global mean estimates on land-ocean data show a gradual increase in yearly temperature anomalies from 1880 until 2020. Looking at the annual mean, one can see there is more change with temperatures

spiking high or low, but when applying Lowess smoothing, it shows a more steady picture of the relative increase in temperature. Both the mean and Lowess smoothing signify the increasing temperature anomalies year over year for general land and ocean data. It is worth noticing that nineteen of the hottest years have occurred since 2000.

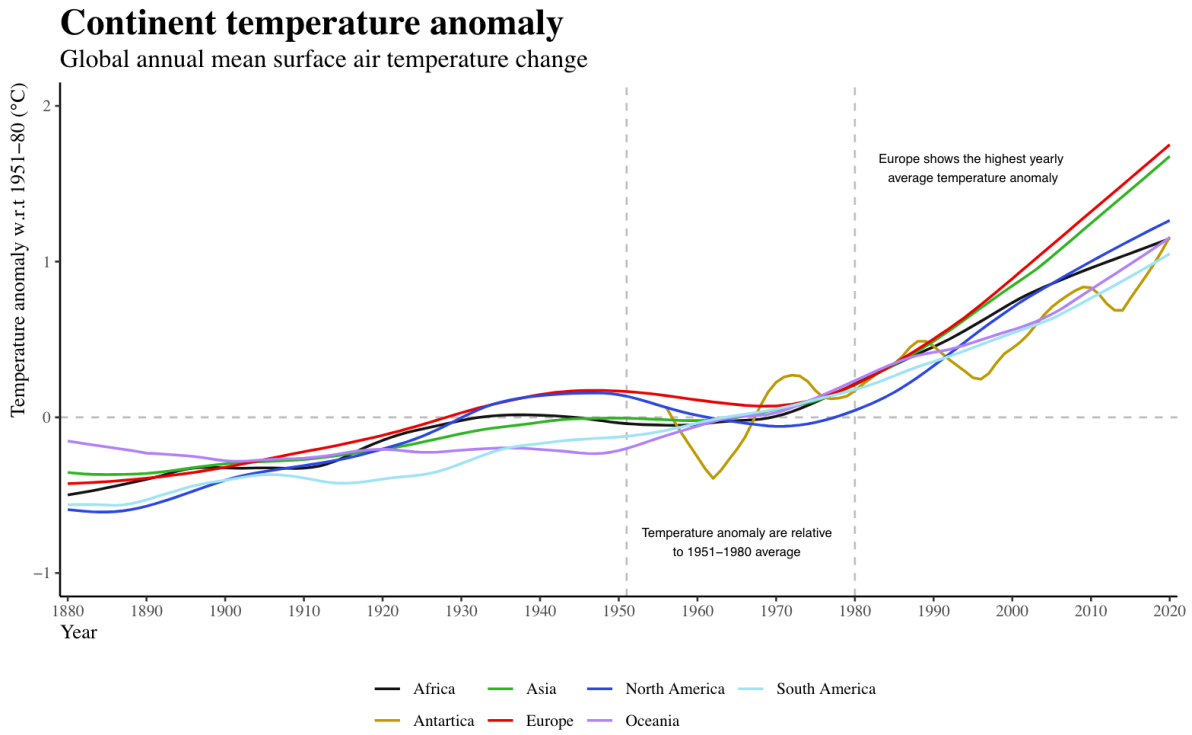

Figure 6. Continent temperature anomaly

In Figure 6, continent anomalies are shown from seven different continents, which all exhibit a gradual increase in temperature from 1880 until 2020. This data is shown on a yearly basis but was aggregated from the raw data Berkeley Earth provided by transforming it from monthly to yearly anomalies. This graph clearly exemplifies how surface air temperatures are increasing year over year for every continent on earth. Six out of seven continents have a steady incline, whereas Antarctica has more variation with years that spike lower or higher, but continues to show a relative increase in surface air temperature. Among all continents, Europe exhibits the highest average temperature anomaly, starting from the mid 1980's.

Lastly, it is interesting to see the years that show the highest average temperature anomaly.

year	yearly_anomaly	lowess	zone
2020	2.163750	1.750654	europe
2020	2.111417	1.676391	asia
2016	1.943500	1.162198	north_america
2019	1.846000	1.707456	europe
2014	1.817583	1.492195	europe
2015	1.802917	1.535166	europe
2018	1.757583	1.664258	europe
2010	1.729000	1.000779	north_america
2019	1.688917	1.632638	asia
2017	1.685833	1.545925	asia

Figure 7. Highest average temperature anomaly per year

The hottest year was registered in 2020 in Europe, with a yearly average temperature anomaly of 2.16 °C relative to the 1951-1980 average. Followed by 2020 in Europe and 2016 in North America. It is worth mentioning that the top ten hottest years per continent were registered after 2014.

The aim of the visualizations was to respond to the first research question. Thus, evidence from two of the most prestigious climate science institutes, NASA/GISS and Berkley Data, shows that the global yearly average temperature, which has risen about 1.18 degrees Celsius in the last century, is increasing at a pace never seen before. More specifically, in the last decade, there were measured the highest temperature anomaly of Earth's surface. Therefore the data support the fact that we are currently experiencing a period of rapid warming of our planet

It can be interesting to visualize the difference per continent during these years. This will exemplify Europe's positioning in comparison, which will be further explored throughout the second research question, when narrowing the paper's scope to European data and what drivers affect climate change.

4. Analysis & Result

4.1 Data Pre-Processing

4.1.1 Data Gathering and Transformation

The Eurostat database was used to gather data concerning the independent variables related to potential drivers influencing climate change. After looking at the three factors to choose the datasets, we applied filtering and transformation techniques to each dataset. More specifically, it was necessary to read each group of data and download it in a tsv.gz format. Only aggregated data of all 28 EU

countries were examined, and the ones related to a singular country were deleted. Then, the dataset was transposed to have all the different variables in the columns and the years in the rows. The correct names for each variable were added to the dataset. Any extra characters (strings, single characters, etc.) in the dataset were eliminated and subsequently all values were transformed to numeric values in order to be processed. For each variable type, the data were merged into one data frame and narrowed down to 56 variables to include data for Greenhouse gas Emissions and drivers: Energy, Transport, Industrial Processes and Products and Waste called df_climate_raw. See appendix B to look at the data gathering and transformation code on R.

4.1.2 Preparation of Dataset for the Model

Afterwards, it is necessary to prepare the data in order to fit the model. Firstly, all the independent variables of the model were merged with the climate anomaly data from NASA/GISS, the dependent variable. In this way a single database containing the dependent variable and independent variables was created. Secondly, data since 1996 has been considered because most of the dataset from Eurostat has 1996 as the first data collection date. Note that not all variables are starting from that date, as an example some variables start from 2000.

	sts_copr_a_PROD_F_CA_I10_EU28	sts_copr_a_PROD_F_CC1_CA_I10_EU28
1996	96.5	NA
1997	95.4	NA
1998	95.7	NA
1999	98.7	NA
2000	101.8	103.7
2001	102.9	104.4
2002	103.4	104.8
2003	105.2	107.0

Figure 8. Final dataset starting years

Therefore, a problem with missing data emerged. Given that the sample size is quite limited, an imputation of missing data was conducted. It is possible to apply several packages solving the problem of imputation of time series data. Most popular and mature are AMELIA, mice and imputeTS. The package imputeTS was used in the data processing part because it specialized on time series imputation and uses time dependencies instead of inter-attribute correlation (Moritz, Bartz-Beielstein, 2017). The package offers multiple algorithm implementations and the missing value imputation by Kalman Smoothing was used because very often produce the best results when trends are presented (Moritz, Bartz-Beielstein, 2017).

As a last step, a correlations analysis among the different variables was conducted. All the variables with correlations higher than 0.9 were rejected in order to avoid multicollinearity problems. Through this process from the 55 starting variables, only 20 were taken into consideration.

4.2 Modeling

The following section will examine regression analysis with the processed data to understand whether and to which extent they influence climate change. See Appendix C to see the R code related to the modeling section.

4.2.1 Base Model

A linear regression model was performed after the data gathering, transformation, and processing steps. 20 variables were considered as independent variables and the climate anomaly is considered as an dependent variable, called "No_smoothing" which is the average annual yearly anomaly temperature.

Dependent variable	Average annual yearly anomaly temperature *No_smoothing*
Number of observation	26
Type	OLS linear regression
Residual standard error:	0.0449 on 5 degrees of freedom
Multiple R^2	0.989
Adjusted R^2	0.9452
F-statistic	22.56 on 20 and 5 DF
F-statistic p-value	0.001329

Figure 9. Base model info and fit

$$F(20, 5) = 22.56, p < 0.01$$

F Statistics indicates that the model is statistically significant using the rule of thumb of p value below (.05).

By looking at the R-squared, the model explains 98.9% of the variance. Since the model has 21 independent variables, it is more informative to look at the adjusted R-squared which penalizes the number of variables in the model. By looking at the adjusted R-squared, the model explains 94.5% of the variance.

	Estimate	Standard Error	Pr(>|t|)
(Intercept)	22.20	8.185	0.04214 *
sts_copr_a_PROD_F_CC1_CA_I10_EU28	$-5.543e-02$	$2.313e-02$	0.06187 .
sts_copr_a_PROD_F_CC2_CA_I10_EU28	$4.587e-02$	$2.812e-02$	0.16379
sts_inpr_a_PROD_C_CA_I10_EU28	$6.138e-02$	$2.190e-02$	0.03789 *
sts_inpr_a_PROD_D_CA_I10_EU28	$-6.904e-02$	$1.782e-02$	0.01170 *
env_wasmun_GEN_KG_HAB_EU28	$-6.972e-03$	$1.014e-02$	0.52215
env_wasmun_TRT_KG_HAB_EU28	$-3.643e-03$	$1.186e-02$	0.77109
tai08_CO2_PC_CRF3_EU28	$1.796e+00$	$1.118e+00$	0.16901
tran_hv_psmod_PC_BUS_TOT_EU28	$-6.441e-02$	$2.843e-01$	0.82975
tran_hv_psmod_PC_TRN_BUS_TOT_AVD_EU28	$-3.288e-01$	$2.370e-01$	0.22404
tran_hv_pstra_I10_EU28	$-3.241e-02$	$8.067e-02$	0.70451
ttr00005_TOT_LOADED_THS_T_EU28	$-5.819e-07$	$2.572e-07$	0.07314 .
t2020_rk200_TOTAL_KTOE_FC_OTH_HH_E_EU28	$2.564e-06$	$2.545e-06$	0.35983
ten00123_FC_E_C0000X0350.0370_KTOE_EU28	$-1.362e-04$	$6.230e-05$	0.08042 .
ten00123_FC_E_C0350.0370_KTOE_EU28	$6.108e-04$	$2.054e-04$	0.03102 *
ten00123_FC_E_E7000_KTOE_EU28	$-1.197e-04$	$4.119e-05$	0.03356 *
ten00123_FC_E_H8000_KTOE_EU28	$2.348e-04$	$5.517e-05$	0.00804 **
ten00123_FC_E_O4000XBIO_KTOE_EU28	$4.972e-05$	$2.732e-05$	0.12841
ten00123_FC_E_S2000_KTOE_EU28	$2.711e-03$	$4.253e-03$	0.55196
ten00123_FC_E_TOTAL_KTOE_EU28	$-1.374e-06$	$4.934e-06$	0.79180
ten00123_FC_E_W6100_6220_KTOE_EU28	$2.168e-04$	$1.997e-04$	0.32715

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Figure 10. Base model coefficients

By looking at the coefficients, the variable *ten00123_FC_E_H8000_KTOE_EU28* is the only one statistically significant at a significance level of 0.01. This variable represents the final energy consumption of heat, in thousand tonnes oil equivalent. This means that by keeping all the other variables constant, a unit change in final energy consumption of heat in one year results in 0.00023° change in yearly anomaly temperature.

The following variables are the ones statistically significant at a significance level of 0.05.

- *sts_copr_a_PROD_F_CC1_CA_I10_EU28 represent*
- *sts_inpr_a_PROD_C_CA_I10_EU28*
- *sts_inpr_a_PROD_D_CA_I10_EU28*
- *ten00123_FC_E_C0350.0370_KTOE_EU28*
- *ten00123_FC_E_E7000_KTOE_EU28*

4.2.2 Diagnostic of Base Model

Diagnostic tests are fundamental to verify whether the results of the model are reliable.

Multicollinearity

	VIF
sts_copr_a_PROD_F_CC1_CA_I10_EU28	294.69597
sts_copr_a_PROD_F_CC2_CA_I10_EU28	175.23350
sts_inpr_a_PROD_C_CA_I10_EU28	405.76854
sts_inpr_a_PROD_D_CA_I10_EU28	117.06705
env_wasmun_GEN_KG_HAB_EU28	343.06186
env_wasmun_TRT_KG_HAB_EU28	319.40429
tai08_CO2_PC_CRF3_EU28	23.80831
tran_hv_psmod_PC_BUS_TOT_EU28	377.59264
tran_hv_psmod_PC_TRN_BUS_TOT_AVD_EU28	126.15657
tran_hv_pstra_I10_EU28	2093.51451
ttr00005_TOT_LOADED_THS_T_EU28	186.04308
t2020_rk200_TOTAL_KTOE_FC_OTH_HH_E_EU28	12.39700
ten00123_FC_E_C0000X0350.0370_KTOE_EU28	222.41453
ten00123_FC_E_C0350.0370_KTOE_EU28	55.56415
ten00123_FC_E_E7000_KTOE_EU28	182.27139
ten00123_FC_E_H8000_KTOE_EU28	60.47841
ten00123_FC_E_O4000XBIO_KTOE_EU28	2830.04005
ten00123_FC_E_S2000_KTOE_EU28	36.18083
ten00123_FC_E_TOTAL_KTOE_EU28	71.89523
ten00123_FC_E_W6100_6220_KTOE_EU28	427.26112

Figure 11. Base model VIF test

The first test run was the VIF test to measure multicollinearity. It evaluates the correlation between multiple variables in the model. Since each VIF value is higher than 10, the model is characterized by multicollinearity, meaning the statistical significance of the model will be off.

Normality of residuals

Shapiro-Wilk Normality Test

data: base model residuals

W	0.91189
p-value	0.02913

Figure 12. Base model Shapiro-Wilk normality test

Subsequently, the Shapiro-Wilk normality test has been conducted to measure the normality of residuals. The low p-value (< 0.05) indicates the residuals are not normally distributed.

Furthermore, theoretical (expected) quantiles and sample (actual) quantiles should be equal or at least very close. Through the following graph, it is possible to visualize that in the model theoretical quantiles and samples do not match, showing again the residuals are not normally distributed.

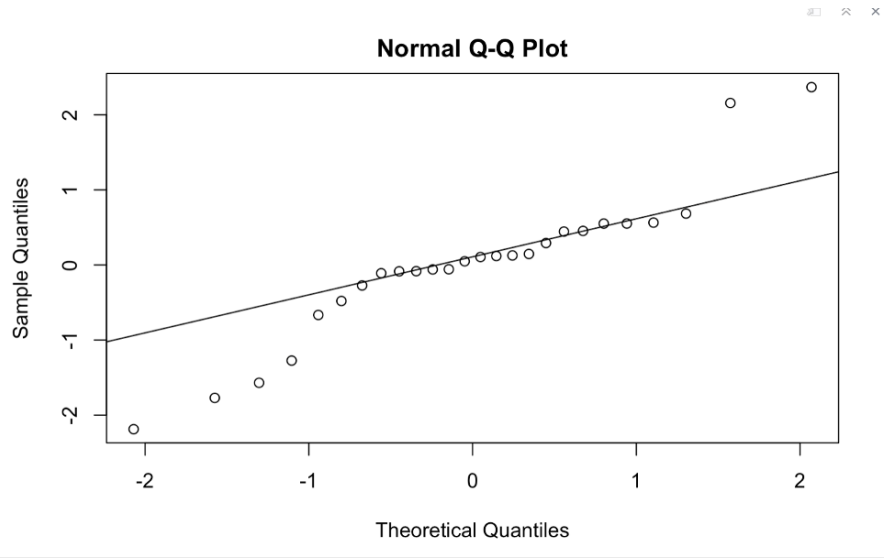

Figure 13. Base model normal Q-Q Plot

Autocorrelation

Durbin-Watson Test

Alternative hypothesis: rho != 0

log	Autocorrelation	D-W Statistic	p-value
1	−0.4822647	2.937878	0.482

Figure 14. Base Durbin Watson Test

Moreover, it is crucial that a regression model is characterized by the independence of residuals. Hence, a Durbin Watson Test has been conducted in order to measure the correlation among the residuals. The results indicate there is no autocorrelation of the residuals since the p-value is higher than 0.05.

Heteroskedasticity

Non-constant Variance Score Test

Variance formula: fitted values

Chisquare	7.343481
Df	1
p-value	0.02913

Figure 15. Base model Non-constant Variance Score Test

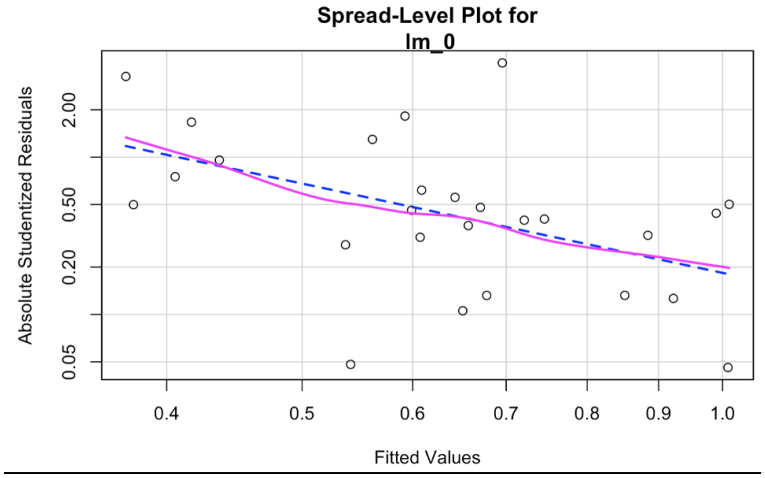

Figure 16. Base model Spread-Level Plot

Lastly, a non-constant variance test was run in order to analyze whether the model presents a heteroskedasticity or a homoskedasticity. The p value is small ($p < 0.01$), hence the null hypothesis of homoscedasticity is rejected. The results indicate that the model is characterized by heteroskedasticity, meaning it is characterized by non-constant variance and that the model is more reliable for certain values of estimated values (where variance is smaller) and less reliable for other values.

In conclusion, the results that emerged by the diagnostic test of the model highlights that the model suffers from multicollinearity, residuals not normally distributed and heteroskedasticity.

4.3 Model Improvement

In this section different techniques to increase the validity of the model were tested, namely imputation of missing data to increase sample size, standardization and stepwise regression.

4.3.1 Increase the Sample Size

In order to solve the aforementioned problems, the sample size has been increased through the imputation of missing data. More specifically, the analysis was conducted again starting from 1980 using the imputation by Kalman Smoothing and State Space Models that estimate the missing values considering the trend of the recovered ones. The model presented the following results:

Dependent variable	Average annual yearly anomaly temperature $No\_smoothing$
Number of observation	36
Type	OLS linear regression (More sample size)
Residual standard error:	0.08998 on 20 degrees of freedom
Multiple R^2	0.9196
Adjusted R^2	0.8593
F-statistic	15.25 on 15 and 20 DF
F-statistic p-value	$8.686e-08$

Figure 17. More sample size model info and fit

$$F(15, 20) = 15.25, p < 0.001$$

F Statistics indicates that the model is statistically significant using the rule of thumb of p-value below (.05) and it is lower than the p-value of the previous linear regression (0.0013).

By looking at the adjusted R-squared, the model explains 85.9% of the variance. Which means that the explanatory power of the model has now decreased, compared to the base model which explained 94% of the variances.

Furthermore, the results of the diagnostic tests for the new model highlighted that the multicollinearity problem has now decreased. More specifically, the VIF test shows 4 values lower than 10, whearease in the base model no variables showed a VIF lower than 10.

Moreover, the Shapiro-Wil normality test shows a p-value higher than 0.5 indicating the residuals are normally distributed, an improvement compared to the base model. This can be easily seen through the following graph.

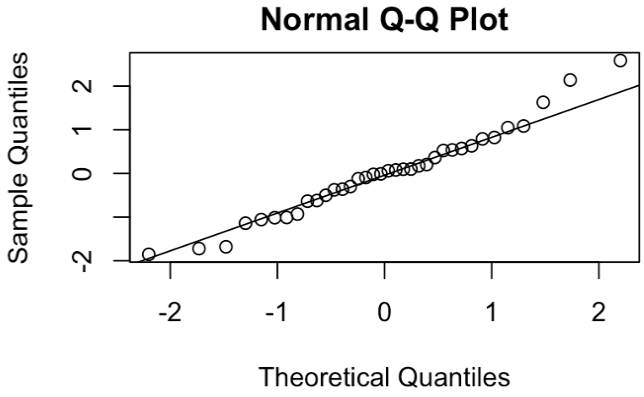

Figure 18. More sample size model Normal Q-Q Plot

Furthermore, the model shows still no autocorrelation among the residuals since the p-value emerged by the Durbin Watson Test is still higher than 0.05. On the other hand, the model persists in showing a heteroskedasticity problem since the results of the NCV test indicates a low p-value, meaning that we still reject the null hypothesis of homoskedasticity.

4.3.2 Standardization

Many independent variables present different orders of magnitude. Thus, a standardization technique was applied to check whether it can improve model performance. Standardization is a commonly used statistical technique that places different variables on an identical scale.

Using the function *scale*, the dataset was standardized. The model emerged presented the following results:

Dependent variable	Average annual yearly anomaly temperature $No\_smoothing$
Number of observation	36
Type	OLS linear regression (Standardize)
Residual standard error:	0.3751 on 20 degrees of freedom
Multiple R^2	0.9196
Adjusted R^2	0.8593
F-statistic	15.25 on 15 and 20 DF
F-statistic p-value	$8.686e-08$

Figure 19. Standardize model info and fit

$$F(15, 20) = 15.25, p < 0.001$$

F Statistics indicates that the model is statistically significant using the rule of thumb of p-value below (.05).

4.3.3 Stepwise Regression

As stated by Denison and Henderson (1989) a stepwise regression can be defined as the set of model comparison and repetitive search operations identifying which are the independent variables with the highest connection with the dependent ones.

Using the function *step,* stepwise regression was applied to the dataset with the estimation of missing value from 1980. The *step* function chooses the best model by Akaike's information criterion (AIC) in a stepwise algorithm. 'Forward' was selected as the direction of the stepwise search. The model emerged presented the following results:

Dependent variable	Average annual yearly anomaly temperature $No\_smoothing$
Number of observation	36
Type	OLS linear regression (Stepwise regression)
Residual standard error:	0.08111 on 33 degrees of freedom
Multiple R^2	0.8922
Adjusted R^2	0.8857
F-statistic	136.6 on 2 and 33 DF
F-statistic p-value	$2.2e-16$

Figure 20. Stepwise regression model info and fit

$$F(2, 33) = 136.6, p < 0.001$$

F Statistics indicates that the model is statistically significant using the rule of thumb of p-value below (.05). By looking at the adjusted R-squared, the model explains 89.2% of the variance. The Stepwise Regression algorithm only considered two variables:

- *sts_inpr_a_PROD_C_CA_I10_EU28*: Volume index relative to 2010 (=100) in manufacture of food products;
- *ten00123_FC_E_S2000_KTOE_EU28* - Final energy consumption of oil shale and oil sands measured as thousand tonnes of oil equivalent.

4.4 Model Evaluation and Comparison

Relative model performance metrics, such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), are used to compare time series models. Those metrics are the best approach when dealing with small data and time series data (Zajic, 2019). When data are recent the split into train, test and validation is not the optimal way to compare models (Zajic, 2019).

AIC is most often used for model selection and can be also used to compare scores of different models. This performance metric is defined as relative, because it gives an overview on how good a model is relative to other models (Akaike, 1974). AIC is defined as:

$$AIC = 2k - 2ln(L)$$

where k is the number of parameters in the model, and L is the maximum value of the likelihood function for the model.

Among the candidate models, the one with the lowest AIC scores is preferred, because it is the one that minimizes the information loss. It is worth noticing that the absolute values of the AIC scores do not matter and that the scores can be negative or positive. In R the function *aictab* was used to calculate AIC.

As stated by Schwarz (1978) another statistical method for the comparative evaluation of time-series models, closely related to the AIC, is the Bayesian Information Criterion (BIC), which is also known as Schwarz Criterion. BIC consists of the estimation of the probability of a model to be true under a given Bayesian setup. Once again, the model to be preferred is the one with the lower BIC value. The BIC value of a model is calculated with the following formula:

$$BIC = k\,ln(n) - 2\,ln(L)$$

Where k is the number of parameters in the model, L is the maximum value of the likelihood function for the model, and n is the number of data points (sample size). In R the function *BIC* was used to calculate AIC.

In this paper AIC and BIC are used to compare the different models used to improve the base one.

Model	# Parameters	AIC	BIC
stepwise_regression	4	-72.53989	-67.49614
more_sample_size	17	-24.38827	-31.46844
standardize	17	78.39691	71.31674

Figure 21. Relative model performance metrics comparison

The *stepwise_regression* model is the one preferred compared to the other, because it has the lowest AIC and BIC scores. The table gives a ranking of the different models based on AIC and BIC criteria, which is a probabilistic ranking of the models that are likely to minimize information loss.

4.4.1 Results of the Preferred Model

From the relative model performance metrics, the preferred model is the one using the stepwise regression.

	Estimate	Standard Error	Pr(>|t|)
(Intercept)	−0.727173	0.145886	$1.93e-05$ ***
sts_inpr_a_PROD_C10_CA_I10_EU28	0.016324	0.001253	$1.46e-14$ ***
ten00123_FC_E_S2000_KTOE_EU28	−0.005231	0.001384	0.000628 ***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\cdot}p < 0.1$

Figure 22. Coefficients of stepwise regression model

The coefficient shows that the variable *sts_inpr_a_PROD_C_CA_I10_EU28* is statistically significant at a level of 0.001. As previously mentioned, this variable measures the manufacture of foods products, related to the 2010 value as an index and there is a positive correlation with the dependent variable, the temperature anomaly change with respect to 1951-1980 average. Thus, given all the other variables constant, a unit change in manufacture of food products will lead to a 0.016°C increase in yearly anomaly temperatures.

Moreover, also the variable *ten00123_FC_E_S2000_KTOE_EU28,* which measures the final energy consumption of oil shale and oil sands, is statistically significant at a significance level of 0.001. In this

case the correlation between this variable and the climate anomaly change is negative. Indeed, by keeping all the other variables constant an increase of one unit in consumption of oil shale and oil sands will result in 0.005°C decrease in yearly anomaly temperature.

4.4.2 Diagnostic of the Preferred Model

This section will investigate the diagnostic tests of the chosen model in order to confirm its significance.

Multicollinearity

	VIF
sts_inpr_a_PROD_C10_CA_I10_EU28	1.212378
ten00123_FC_E_S2000_KTOE_EU28	1.212378

Figure 23. Coefficients of stepwise regression model

The model is not experiencing multicollinearity issues, since the VIF value of the variables is significantly lower than 10. This means there are no sizable correlations between multiple variables within the model.

Normality of residuals

Shapiro-Wilk Normality Test
data: Stepwise regression residuals

W	0.98429
p-value	0.8779

Figure 24. Coefficients of stepwise regression model

As it is also possible to see by the following graph, the residuals of the model are normally distributed, since the p-value=0.87 is considerably high (higher than 0.05).

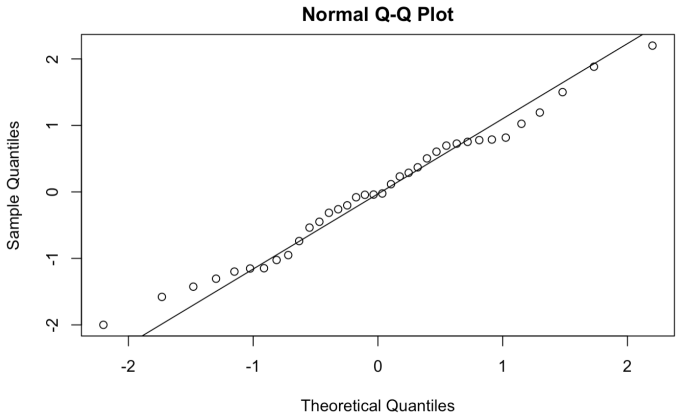

Figure 25. Stepwise regression model Normal Q-Q Plot

Autocorrelation

Durbin-Watson Test

Alternative hypothesis: rho != 0

log	Autocorrelation	D-W Statistic	p-value
1	0.1164049	1.639169	0.118

Figure 26. Stepwise regression model Durbin-Watson Test

In this model there is no significant relation between the residuals, the error terms are instead independent. Thus, there is no suspect of autocorrelation given that the p-value is higher than 0.05. This means that the significance of the test is not impacted.

Homoscedasticity or Heteroscedasticity

Non-constant Variance Score Test

Variance formula: fitted values

Chisquare	5.036168
Df	1
p-value	0.024823

Figure 26. Stepwise regression model Durbin-Watson Test

By conducting an NCV test it is possible to examine whether the model presents homoscedasticity or heteroscedasticity. Since the p-value is lower than 0.05, the null hypothesis of homoscedasticity is rejected. This means that the model does suffer from non-constant variance.

5. Conclusions

5.1 Analysis of the Research Question

After using Berkeley Earth, NASA/GISS and Eurostat data, it can be said that there are two key findings. In relation to the first research question, evidence from different sources shows that yearly average temperatures are increasing at an incredible pace in the last decade. In relation to the second research question, the stepwise regression model shows that the main drivers of climate change are manufacturing of food products and final energy consumption of oil shale & oil sands. From the analysis performed it is possible to conclude that one key driver for climate change and increase in yearly anomaly temperature is production of food. This can be seen because "the food system contributes carbon dioxide emissions that emanate from the use of fossil fuels in transportation, processing, retailing, storage, and preparation" (Carlsson-Kanyama et al., 2009). Food consumption takes into account various factors that affect the larger ecological system, which explains why the increase of such consumption affects the climate. Now that this driver has been identified, further research has recommended that changes in the diet toward more plant-based foods, toward meat from animals with little enteric fermentation, and toward foods processed in an energy-efficient manner can offer opportunities to mitigate climate change (ibid).

A surprising finding was the driver of consumption of oil shale and oil sands. One might think that oil would increase climate change but the effect is actually the opposite, according to our results. While there is a minimal decrease in yearly anomaly temperature, it is a decrease nonetheless. This is contrary to most research, but what is clear is that the oil industry is a vital part of the solution in the transition to a lower carbon energy future and it is up to governments to set a framework of carbon policy and regulation, for temperature anomalies to lower and climate change to be slowed (Lovell, 2011).

5.2 Limitations and Future Research

Given the small size of the datasets used for the modeling process, imputation of missing data was used within this study through the Kalman Smoothing Algorithm. It is crucial to underline that, although the technique is based on the statistical estimation of the trend of time series data, the missing values obtained are still an estimation. For this reason this method should be avoided in real scenarios, since the predicted values can be different from the actual ones.

Furthermore, this study only uses a Linear Regression approach. Future research can increase the model complexity by considering Logit Regression or Neural Network.

6. References

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716-723.

Buis A. (2021). Milankovitch (Orbital) Cycles and Their Role in Earth's Climate. climate.nasa.gov. NASA's Jet Propulsion Laboratory.

Carlsson-Kanyama, A., González, A. (2009). Potential contributions of food consumption patterns to climate change. The American Journal of Clinical Nutrition, Volume 89, Issue 5, Pages 1704S–1709S, https://doi.org/10.3945/ajcn.2009.26736AA

Clayton, L., Attig, J. W., Mickelson, D. M., Johnson, M. D., & Syverson, K. M. (2006). Glaciation of Wisconsin. Madison^ eWisconsin Wisconsin: Wisconsin Geological and Natural History Survey.

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R., Verheggen, B., Maibach, E. W., ... & Rice, K. (2016). Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. Environmental Research Letters, 11(4), 048002.

Gaffney, O., Steffen, W. (2017). The Anthropocene equation. The Anthropocene Review, 4(1), 53-61.

GISTEMP Team. (2021). GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies. https://data.giss.nasa.gov/gistemp/.

Henderson, D. A., & Denison, D. R. (1989). Stepwise regression in social and psychological research. Psychological Reports, 64(1), 251-257.

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. R J., 9(1), 207.

Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss. (2019). Improvements in the GISTEMP uncertainty model. J. Geophys. Res. Atmos., 124, no. 12, 6307-6326, doi:10.1029/2018JD029522.

Lindsey, R. (2009). Climate and Earth's Energy Budget. NASA Earth Observatory.

The Intergovernmental Panel on Climate Change (IPCC). (2021) Sixth Assessment Report, Summary for Policymakers. https://www.ipcc.ch/report/ar6/wg1/#SPM

Lovell, B. (2011). Challenged by carbon: The oil industry and climate change. Cambridge: Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.

Wuebbles, D. J., Fahey, D. W., Hibbard, K. A., Arnold, J. R., DeAngelo, B., Doherty, S., ... & Walsh, J. (2017). Climate science special report: Fourth national climate assessment (NCA4), Volume I.

Zajic, A. (2019). Introduction to AIC - Akaike Information Criterion. Towards Data Science.

7. Appendix

Appendix A

R Code related to climate change evidence

Find the R code named *1_cliamte_change_evidence* in the attached zip file. Both .pdf and .Rmd file are available.

Appendix B

R Code related to data gathering and transformation

Find the R code named *2_data_gathering_transformation* in the attached zip file. Both .pdf and .Rmd file are available.

Appendix C

R code related to the modeling used

Find the R code named *3_modeling* in the attached zip file. Both .pdf and .Rmd file are available.

Appendix D

Detailed dataset description from Eurostat

1. **GREENHOUSE GAS EMISSION**

Name of the dataset: Greenhouse gas emissions by sector (sdg_13_10)
Availability: from 1990 to 2019 (?)
Geography: for every European country
Frequency: Yearly data
Unit of measure: Index 1990 = 100 and tonnes of CO2 equivalent per capita
Use in analysis: Yes
Data description: The indicator measures total national emissions (from both ESD and ETS sectors) including international aviation of the so called 'Kyoto basket' of greenhouse gases, including carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), and the so-called F-gases (hydrofluorocarbons, perfluorocarbons, nitrogen triflouride (NF3) and sulphur hexafluoride (SF6)) from all sectors of the GHG emission inventories (including international aviation and indirect CO2). The indicator is presented in two forms: as net emissions including land use, land use change and forestry (LULUCF) as well as excluding LULUCF. Using each gas' individual global warming potential (GWP), they are being integrated into a single

indicator expressed in units of CO2 equivalents. The GHG emission inventories are submitted annually by the EU Member States to the United Nations Framework Convention on Climate Change (UNFCCC).
The average population of the reference year (calculated as the arithmetic mean of the population on 1st January of two consecutive years) is used as denominator (per capita).

Name of the dataset: Air emissions accounts by NACE Rev. 2 activity (env_ac_ainah_r2)
Availability: 2010-2019
Geography: Every European country
Frequency: yearly
Unit of measure: Index 1990 = 100 and tonnes of CO2 equivalent per capita
Use in analysis: No
Data description: The indicator measures total national emissions (from both ESD and ETS sectors) including international aviation of the so called 'Kyoto basket' of greenhouse gases, including carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), and the so-called F-gases (hydrofluorocarbons, perfluorocarbons, nitrogen triflouride (NF3) and sulphur hexafluoride (SF6)) from all sectors of the GHG emission inventories (including international aviation and indirect CO2). The indicator is presented in two forms: as net emissions including land use, land use change and forestry (LULUCF) as well as excluding LULUCF. Using each gas' individual global warming potential (GWP), they are being integrated into a single indicator expressed in units of CO2 equivalents. The GHG emission inventories are submitted annually by the EU Member States to the United Nations Framework Convention on Climate Change (UNFCCC). The average population of the reference year (calculated as the arithmetic mean of the population on 1st January of two consecutive years) is used as denominator (per capita).

Name of the dataset: Air emissions accounts totals bridging to emission inventory totals (env_ac_aibrid_r2)
Availability: 2010-2019
Geography: Europe
Frequency: yearly
Unit of measure: The air emissions [env_ac_ainah_r2] and the bridging items [env_ac_aibrid_r2] are presented in tonnes and thousand tonnes, as well as grams per capita and kilograms per capita.
Air emissions intensities [env_ac_aeint_r2] are presented in grams per euro and kilograms per euro.
Use in analysis: No
Data description: Air emissions are relevant for monitoring the interaction between the economy and the environment, in particular in a context of global climate change. Air emission accounts data are also used in modelling, including carbon footprint.

The relevance of air emission accounts is enhanced by using a conceptual framework consistent with National Accounts, which allows, e.g. to put in relation with estimates of production, value added, employment, GDP, etc.
The users include policy makers in environmental ministries, environmental organisations, students and interested citizens.
The policy context is set here: http://ec.europa.eu/eurostat/web/environment/overview/policy-context

Name of the dataset: Air emissions intensities by NACE Rev. 2 activity (env_ac_aeint_r2)
Availability: 2010-2019
Geography: Europe

Frequency: yearly
Unit of measure: The air emissions [env_ac_ainah_r2] and the bridging items [env_ac_aibrid_r2] are presented in tonnes and thousand tonnes, as well as grams per capita and kilograms per capita.
Air emissions intensities [env_ac_aeint_r2] are presented in grams per euro and kilograms per euro.
Use in analysis: No
Data description: Air emissions are relevant for monitoring the interaction between the economy and the environment, in particular in a context of global climate change. Air emission accounts data are also used in modelling, including carbon footprint.
The relevance of air emission accounts is enhanced by using a conceptual framework consistent with National Accounts, which allows, e.g. to put in relation with estimates of production, value added, employment, GDP, etc.

Name of the dataset: Emissions of greenhouse gases and air pollutants from final use of CPA08 products - input-output analysis, ESA 2010 (env_ac_io10)
Availability: 2010-2019
Geography: EU economy
Frequency: yearly
Unit of measure: The air emissions 'embodied' in products for final use are presented in tonnes, thousand tonnes and kilograms per inhabitant. Several of the air pollutants are also expressed in equivalents of another air pollutant.
Use in analysis: No
Data description: Air emissions caused by final use of CPA products provide information on the proximate causes of air emissions, by linking air emissions to final use of products.
The users include policy makers in environmental ministries, environmental organisations, journalists, students, and interested citizens.

Name of the dataset: Greenhouse gas emissions in ESD sectors (t2020_35)
Availability: 2011-2019
Geography: Europe
Frequency: yearly
Unit of measure: million tonnes CO2 equivalent and EU Effort Sharing Decision base year = 100
Use in analysis: No
Data description: Indicator is one of the headline indicators of the EU 2020 strategy. It is used to monitor progress towards the EU's target of 'reducing the GHG emissions by at least 20 % compared with 1990 levels' by 2020.
The Energy Union supports the shift towards a resource-efficient, low-carbon economy to achieve sustainable growth through their legal frameworks and related initiatives (see above under sustainable development goal no 7). Relevant legislations have been proposed to support these policies. Most importantly, the European Council has agreed on three key targets for the year 2030: at least 40 % cuts in greenhouse gas emissions (from 1990 levels), at least 27 % share for renewable energy and at least 27 % improvement in energy efficiency. The European Commission has proposed to increase the energy efficiency target to 30 %.
With transport being one of the key sectors to meet the EU’s commitments under the Paris agreement, the European Strategy for Low-Emission Mobility makes an important contribution to reducing GHG emissions

in this sector. Furthermore, the EU plans on Accelerating Clean Energy Innovation to facilitate the clean energy transition through targeted research and innovation.

Name of the dataset: Greenhouse gas emissions from agriculture (tai08)
Availability: 2008-2019
Geography: Europe
Frequency: yearly
Unit of measure: percentage
Use in analysis: Yes
Data description: This indicator tracks trends in greenhouse gas (GHG) emissions by agriculture, estimated and reported under the United Nations Framework Convention on Climate Change (UNFCCC), the Kyoto Protocol and the Decision 525/2013/EC.

Name of the dataset: Greenhouse gas emissions intensity of energy consumption (source: EEA and Eurostat) (sdg_13_20)
Availability: 2000-2019
Geography: Europe
Frequency: yearly
Unit of measure: index 2000 = 100
Use in analysis: No
Data description: The indicator is part of the EU Sustainable Development Goals (SDG) indicator set. It is used to monitor progress towards SDG 13 on climate action and SDG 7 on affordable and clean energy; which are embedded in the European Commission's Priorities under the European Green Deal.
SDG 13 aims to implement the commitment to the United Nations Framework Convention on Climate Change and operationalise the Green Climate Fund. It aims to strengthen countries' resilience and adaptive capacity to climate-related hazards and natural disasters by integrating climate change mitigation and adaptation measures into national strategies, policies and planning. SDG 7 calls for ensuring universal access to modern energy services, improving energy efficiency and increasing the share of renewable energy.

Name of the dataset: Average CO2 emissions per km from new passenger cars (source: EEA, DG CLIMA) (sdg_12_30)
Availability: 2000-2019
Geography: Europe
Frequency: yearly
Unit of measure: g CO2 per km
Use in analysis: Yes
Data description: The indicator is part of the EU Sustainable Development Goals (SDG) indicator set. It is used to monitor progress towards SDG 12 on ensuring sustainable consumption and production patterns, SDG 9 on building resilient infrastructure, promoting inclusive and sustainable industrialisation and fostering innovation, and on SDG 13 on taking urgent action to combat climate change and its impacts. These SDGs are embedded in the European Commission's Priorities under the 'European Green Deal', 'An economy that works for people', and 'A Europe fit for the digital age'.

2. DRIVERS - TRANSPORT

Name of the dataset: Air transport of passengers by country (yearly data) (ttr00012)
Availability: 2009 - 2019
Geography: for every European country
Frequency: Yearly
Unit of measure: Number of passengers carried
Use in analysis: Yes
Data description: The Air transport domain contains national and international intra and extra-EU data. This provides air transport data for passengers.

Name of the dataset: Air transport of goods by country (yearly data) (ttr00011)
Availability: 2009 - 2019
Geography: for every European country
Frequency: Yearly
Unit of measure: Tonne of Freight and mail loaded and unloaded
Use in analysis: Yes
Data description: The Air transport domain contains national and international intra and extra-EU data. This provides air transport data for tonne of Freight and mail loaded and unloaded.

Name of the dataset: Sea transport of goods (ttr00009)
Availability: 2009-2019
Geography: for every European country
Frequency: Yearly
Unit of measure: Thousand tonnes
Use in analysis: Yes
Data description: The table displays the gross weight of seaborne goods handled in ports (goods unloaded from vessels plus goods loaded onto vessels)

Name of the dataset: Goods transport by rail (ttr00006)
Availability: 2011-2016
Geography: for every European country
Frequency: Yearly
Unit of measure:
Use in analysis: No → few data
Data description: Data displayed in this table cover the Rail transport of goods which relate Rail goods transport in the Member States on its national territory.

Name of the dataset: Air passenger transport between main airports in each reporting country and partner reporting countries (avia_paoac)
Availability: 1993 - 2020
Geography: for every European country
Frequency: Yearly, quarterly, monthly
Unit of measure: Passengers on board, passengers carried, commercial passengers air flight
Use in analysis: No → Data from every airport, need more general

Data description: The Air transport domain contains national and international intra and extra-EU data. This provides air transport data for passengers (in number of passengers) and for freight and mail (in 1 000 tonnes) as well as air traffic data by airports, airlines and aircraft. The air transport data is collected at airport level. As from 2003 reference year the data are provided according to the legal act (some countries were given derogation until 2005). Until 2002 partial information (passenger transport only) was available for some countries and airports.

Name of the dataset: Freight and mail air transport between main airports in each reporting country and partner reporting countries (avia_gooac)
Availability: 1993 - 2020
Geography: for every European country
Frequency: Yearly, quarterly, monthly
Unit of measure: freight and mail on board, freight and mail loaded, commercial passengers air flight
Use in analysis: Too much data processing
Data description: The Air transport domain contains national and international intra and extra-EU data. This provides air transport data for passengers (in number of passengers) and for freight and mail (in 1 000 tonnes) as well as air traffic data by airports, airlines and aircraft. The air transport data is collected at airport level. As from 2003 reference year the data are provided according to the legal act (some countries were given derogation until 2005). Until 2002 partial information (passenger transport only) was available for some countries and airports.

Name of the dataset: Summary of annual road freight transport by type of operation and type of transport (1 000 t, Mio Tkm, Mio Veh-km) (road_go_ta_tott)
Availability: 1999 - 2020
Geography: ONLY for for every European country
Frequency: Yearly
Unit of measure: type of carriage, type of operation status
Use in analysis: No → data not availble for europe in total
Data description: Eurostat collects road transport statistics by two means:
1. Data on infrastructure, transport equipment, enterprises, economic performance, employment, traffic, aggregated data on transport of passengers and goods as well as data on accidents are collected using the Common Questionnaire of the United Nations Economic Commission for Europe (UNECE), Eurostat and the International Transport Forum (ITF, in the framework of OECD). The method of the Common Questionnaire data collection is presented in a separate document.
2. Data on carriage of goods by road, using heavy goods vehicles, are based on a continuum of legal acts:

Name of the dataset: Modal split of freight transport (tran_hv_frmod)
Availability: 2005 - 2020
Geography: for every European country
Frequency: Yearly
Unit of measure: Percentage of mode of transport
Use in analysis: No → few data
Data description: This indicator is defined as the percentage of each inland mode in total freight transport performance measured in tonne-kilometres. Inland freight transport modes include road, rail and inland waterways.

Name of the dataset: Volume of freight transport relative to GDP (tran_hv_frtra)
Availability: 2005 - 2020
Geography: for every European country
Frequency: Yearly
Unit of measure: Index of 2010
Use in analysis: Yes
Data description: Index of inland freight transport volume relative to GDP, 2010=100.
This indicator is defined as the ratio between tonne-kilometres (inland modes) and GDP (chain-linked volumes, at 2010 exchange rates). It is indexed on 2010.
Inland freight transport includes road, rail and inland waterways.

Name of the dataset: Modal split of passenger transport (tran_hv_psmod)
Availability: 1990 - 2020
Geography: for every European country
Frequency: Yearly
Unit of measure: Percentage of volume
Use in analysis: Yes
Data description: This indicator is defined as the percentage of transport by passenger cars, buses and coaches, and trains in total inland passenger transport performance, measured in passenger-km.
Inland passenger transport includes road (passenger cars, buses and coaches) and rail (trains) transport.

Name of the dataset: Volume of passenger transport relative to GDP (tran_hv_pstra)
Availability: 1990 - 2020
Geography: for every European country
Frequency: Yearly
Unit of measure: Volume of passenger. Indexed at 2010
Use in analysis: Yes
Data description: This indicator is defined as the ratio between the total transport performance of passengers using the inland modes (road and rail), expressed in passenger-kilometres and GDP (chain-linked volumes, at 2010 exchange rates).
It is indexed 2010=100.
Total inland passenger transport includes road transport (transport by passenger cars and buses/coaches) and rail transport (by trains).

Name of the dataset: Final energy consumption in transport by type of fuel (ten00126)
Availability: 2009
Geography: for every European country
Frequency: Yearly
Unit of measure: Thousands of tonnes
Use in analysis: No → not enough data (time)
Data description: The energy balance is the most complete statistical accounting of energy products and their flow in the economy. The energy balance allows users to see the total amount of energy extracted from the environment, traded, transformed and used by different types of end-users. It also allows seeing the relative contribution of each energy carrier (fuel, product). The energy balance allows studying the overall

domestic energy market and monitoring impacts of energy policies. The energy balance offers a complete view on the energy situation of a country in a compact format, such as on energy consumption of the whole economy and of individual sectors. The energy balance presents all statistically significant energy products (fuels) of a country and their production, transformation and consumption by different type of economic actors (industry, transport, etc.). Therefore, an energy balance is the natural starting point to study the energy sector.

Name of the dataset: Final energy consumption in transport by type of fuel (ten00126)
Availability: 2009
Geography: for every European country
Frequency: Yearly
Unit of measure: Thousands of tonnes
Use in analysis: No → not enough data (time)
Data description: The energy balance is the most complete statistical accounting of energy products and their flow in the economy. The energy balance allows users to see the total amount of energy extracted from the environment, traded, transformed and used by different types of end-users. It also allows seeing the relative contribution of each energy carrier (fuel, product). The energy balance allows studying the overall domestic energy market and monitoring impacts of energy policies. The energy balance offers a complete view on the energy situation of a country in a compact format, such as on energy consumption of the whole economy and of individual sectors. The energy balance presents all statistically significant energy products (fuels) of a country and their production, transformation and consumption by different type of economic actors (industry, transport, etc.). Therefore, an energy balance is the natural starting point to study the energy sector.

Name of the dataset: Final energy consumption in road transport by type of fuel (ten00127)
Availability: 2009
Geography: for every European country
Frequency: Yearly
Unit of measure: Thousands of tonnes
Use in analysis: No → not enough data (time)
Data description: The energy balance is the most complete statistical accounting of energy products and their flow in the economy. The energy balance allows users to see the total amount of energy extracted from the environment, traded, transformed and used by different types of end-users. It also allows seeing the relative contribution of each energy carrier (fuel, product). The energy balance allows studying the overall domestic energy market and monitoring impacts of energy policies. The energy balance offers a complete view on the energy situation of a country in a compact format, such as on energy consumption of the whole economy and of individual sectors. The energy balance presents all statistically significant energy products (fuels) of a country and their production, transformation and consumption by different type of economic actors (industry, transport, etc.). Therefore, an energy balance is the natural starting point to study the energy sector.

3. DRIVERS - INDUSTRIAL PROCESS AND PRODUCTS

Name of the dataset: Production in industry - annual data (sts_inpr_a)

Availability: 1980
Geography: for every European country
Frequency: Yearly
Unit of measure: Indices = 2010
Use in analysis: Yes
Data description: Short-term statistics (STS) give information on a wide range of economic activities according to **NACE Rev.2** classification (Statistical Classification of Economic Activities in the European Community). The industrial import price indices offer information according to the **CPA** classification (Statistical Classification of Products by Activity in the European Economic Community). Construction indices are broken down by Classification of Types of Constructions (CC).

All data under this heading are **index** data. **Percentage changes** are also available for each indicator: Infra-annual percentage changes - changes between two consecutive months or quarters - are calculated on the basis of non-adjusted data (prices) or seasonally adjusted data (value and volume indicators) and year-on-year changes - comparing a period to the same period one year ago - are calculated on the basis of non-adjusted data (prices and employment) or calendar adjusted data (volume and value indicators).

Name of the dataset: Production in construction - annual data (sts_copr_a)
Availability: 1990
Geography: for every European country
Frequency: Yearly
Unit of measure: Thousands of tonnes
Use in analysis: Yes
Data description: Short-term statistics (STS) give information on a wide range of economic activities according to **NACE Rev.2** classification (Statistical Classification of Economic Activities in the European Community). The industrial import price indices offer information according to the **CPA** classification (Statistical Classification of Products by Activity in the European Economic Community). Construction indices are broken down by Classification of Types of Constructions (CC).

All data under this heading are **index** data. **Percentage changes** are also available for each indicator: Infra-annual percentage changes - changes between two consecutive months or quarters - are calculated on the basis of non-adjusted data (prices) or seasonally adjusted data (value and volume indicators) and year-on-year changes - comparing a period to the same period one year ago - are calculated on the basis of non-adjusted data (prices and employment) or calendar adjusted data (volume and value indicators).

4. DRIVERS - ENERGY

Name of the dataset: Primary energy consumption (sdg_07_10)
Availability: from 2000 to 2019
Geography: every European country
Frequency: yearly data
Unit of measures: million tonnes of oil equivalent (TOE), index 2005 = 100 and TOE per capita
Use in analysis: Yes
Data description: The indicator measures the total energy needs of a country excluding all non-energy use of energy carriers (e.g. natural gas used not for combustion but for producing chemicals).

"Primary Energy Consumption" covers the energy consumption by end users such as industry, transport, households, services and agriculture, plus energy consumption of the energy sector itself for production and transformation of energies, losses occurring during the transformation of energies (e.g. the efficiency of electricity production from combustible fuels) and the transmission and distribution losses of energy).

Name of the dataset: Final energy consumption
Availability: from 2000 to 2019
Geography: every European country
Frequency: yearly data
Unit of measures: million tonnes of oil equivalent (TOE), index 2005 = 100 and TOE per capita
Use in analysis: Yes
Data description: The indicator measures the energy end-use in a country excluding all non-energy use of energy carriers (e.g. natural gas used not for combustion but for producing chemicals).
"Final energy consumption" covers only the energy consumed by end users, such as industry, transport, households, services and agriculture; it excludes energy consumption of the energy sector itself and losses occurring during transformation and distribution of energy.

Name of the dataset: final energy consumption in households
Availability: from 1990 to 2020
Geography: every European country
Frequency: yearly data
Unit of measures: Final energy consumption in households: Thousand tonnes of oil equivalent (TOE).
- Final energy consumption in households by fuel: Percentage of final energy consumption in households. (can't find the % in the dataset)
Use in analysis: Yes
Data description: The indicator final energy consumption in households measures the total energy consumed by households as final users, expressed in 1000 tonnes of oil equivalent.
As regards final energy consumption in households by fuel, seven fuel types have been taken into account: solid fossil fuels, other fuels, oil and petroleum products, natural gas, electricity, heat and renewables and biofuels.
The indicator has been chosen as a proxy for indicators in the key area 'Improving buildings' of the resource efficiency initiative. This area focuses on the energy spent in households for heating purposes and how the amelioration of buildings can contribute to energy-saving plans.
Eurostat collects data on total energy consumption in households split by fuel category. More detailed data for energy consumption in households (e.g. energy for space heating, space cooling, water heating and cooking) will be collected in the future under the Commission Regulation (EU) No 431/2014 of 24 April 2014 amending Regulation (EC) No 1099/2008 of the European Parliament and of the Council on energy statistics, as regards the implementation of annual statistics on energy consumption in households.

Name of the dataset: Final energy consumption in households per capita
Availability: from 2000 to 2020
Geography: every European country

Frequency: yearly data
Unit of measures: kg of oil equivalent.
Use in analysis: No (similar to the previous one)
Data description: The indicator measures how much energy every citizen consumes at home excluding energy used for transportation. Since the indicator refers to final energy consumption, only energy used by end consumers is considered. The related consumption of the energy sector itself is excluded.

Name of the dataset: final energy consumption by product
Availability: from 2011 to 2020
Geography: 28 European countries
Frequency: yearly
Unit of measures: Thousand tonnes of oil equivalent [KTOE]
Use in analysis: Yes
Data description: Final energy consumption covers the energy consumption of end-users, such as industry, transport, households, services and agriculture. It excludes consumption of the energy sector itself and losses occurring during transformation and distribution of energy (e.g. power plants, district heating plants, oil refineries, coke ovens, blast furnaces). It is also excluding all non-energy use of energy carriers (e.g. natural gas used for producing chemicals, oil based lubricants, bitumen used for road surface). Quantities delivered to international aviation and international marine bunkers are also excluded from the final energy consumption.
– BY PRODUCT (e.g. heat, natural gas, electricity etc.)

Name of the dataset: final energy consumption by sector
Availability: from 2011 to 2020
Geography: 28 european countries
Frequency: Yearly data
Unit of measures: Thousand tonnes of oil equivalent [KTOE]
Use in analysis: No (not so many data)
Data description: Final energy consumption covers the energy consumption of end-users, such as industry, transport, households, services and agriculture. It excludes consumption of the energy sector itself and losses occurring during transformation and distribution of energy (e.g. power plants, district heating plants, oil refineries, coke ovens, blast furnaces). It is also excluding all non-energy use of energy carriers (e.g. natural gas used for producing chemicals, oil based lubricants, bitumen used for road surface). Quantities delivered to international aviation and international marine bunkers are also excluded from the final energy consumption.
– BY SECTOR (e.g. industry, transport)

Name of the dataset: final energy consumption in services by type of fuel
Availability: from 2009 to 2020
Geography: 28 european countries
Frequency: yearly
Unit of measures: Thousand tonnes of oil equivalent [KTOE]
Use in analysis: No
Data description: Final energy consumption in services covers the energy consumption of public and private entities in the NACE divisions 33, 36, 37, 38, 39, 45, 46, 47, 52, 53, 55, 56, 58, 59, 60, 61, 62, 63, 64, 65, 66,

68, 69, 70, 71, 72, 73, 74, 75, 77, 78, 79, 80, 81, 82, 84 (excluding Class 8422), 85, 86, 87, 88, 90, 91, 92, 93, 94, 95, 96 and 99. Also includes fuel used by all non-transport activities of NACE Divisions 49, 50 and 51 (such as heating and lighting of buildings).

Name of the dataset: Final energy consumption in industry by type of fuel
Availability: 2009 to 2020
Geography: 28 countries
Frequency: yearly
Unit of measures: Thousand tonnes of oil equivalent [KTOE]
Use in analysis: No
Data description: Final energy consumption in industry covers the energy consumption of the NACE divisions: 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 41, 42 and 43. For NACE divisions 7, 8 and 9, only mining and quarrying of non-energy products is included. Quantities of energies transformed into another energy product are excluded from the industry sector and reported in the transformation sector (for example: electricity generation, coke ovens, blast furnaces, oil refineries).

Name of the dataset: Complete energy balance
Availability: 2011 - 2020
Geography: 28 european countries
Frequency: yearly
Unit of measures: Thousand tonnes of oil equivalent [KTOE]
Use in analysis: No
Data description: The energy balance is the most complete statistical accounting of energy products and their flow in the economy. The energy balance allows users to see the total amount of energy extracted from the environment, traded, transformed and used by different types of end-users. It also allows seeing the relative contribution of each energy carrier (fuel, product). The energy balance allows studying the overall domestic energy market and monitoring impacts of energy policies.

Name of the dataset: Share of fossils fuels in gross available energy
Availability: 2010 - 2019
Geography: 28 european countries
Frequency: yearly
Unit of measures: %
Use in analysis: No
Data description: quantity of fossils fuels in gross available energy

Name of the dataset: share of fuels in final energy consumption
Availability: 2010-2019
Geography: 28 european countries
Frequency: yearly
Unit of measures: %
Use in analysis: No
Data description: quantity of fuel used in final energy consumption

Name of the dataset: production of electricity and derived heat by type of fuel

Availability: 2010-2019
Geography: 28 countries
Frequency: yearly
Unit of measures: Thousand tonnes of oil equivalent [KTOE]
Use in analysis: No
Data description:

Name of the dataset: Electricity production capacities by main fuel groups and operator
Availability: 2010 to 2019
Geography: 28 eu countries
Frequency: yearly
Unit of measures: Megawatt (MW)
Use in analysis: no
Data description: These are data collected via the annual electricity and heat questionnaire and the annual renewables questionnaire, according to Annex B of the Regulation (EC) No 1099/2008 of the European Parliament and of the Council of 22 October 2008 on energy statistics. The variables are: Total capacity (MWe) Capacity by source of electricity production (MWe) Capacity by type of generation in power plants using combustible fuels (MWe) Capacity by type of firing and by type of fuel used in power plants using combustible fuels (MWe) All reported capacities are broken down by type of supplier (main activity producer or auto-producer) in nrg_inf_epc. For plants based on combustion of fuels the capacity is further divided by type of technology of the generating plant (steam, internal combustion….) in nrg_inf_epct, by type of firing and by type of fuels.

Name of the dataset: Key indicators of physical energy flow accounts by NACE Rev. 2 activity
Availability: 2010-2019
Geography: 28 eu countries
Frequency: yearly
Unit of measures: terajoule (TJ)
Use in analysis: no
Data description: Physical energy flow accounts (PEFA) is one module of the European environmental-economic accounts - Regulation (EU) 691/2011 Annex VI. PEFA record the flows of energy (in terajoules) from the environment to the economy (natural inputs), within the economy (products), and from the economy back to the environment (residuals), using the accounting framework of physical supply and use tables. PEFA provide information on energy flows arranged in a way fully compatible with concepts, principles, and classifications of national accounts – thus enabling integrated analyses of environmental, energy and economic issues e.g. through environmental-economic modelling. PEFA complement the traditional energy statistics, balances and derived indicators which are the main reference data source for EU energy policies. This metadata refers to three PEFA datasets based on the same data collection: Energy supply and use by NACE Rev. 2 activity (env_ac_pefasu), containing data on supply (table A), use (table B), transformation use (table B1), end use (table B2) and emission-relevant use (table C) Key indicators of physical energy flow accounts by NACE Rev. 2 activity (env_ac_pefa04) Physical energy flow accounts totals bridging to energy balances totals (env_ac_pefa05) – KEY INDICATORS SUCH AS Extraction of natural energy inputs [NEI_EXT] Domestic production of energy products [EPRD_DOM] Intermediate consumption of energy products [EPRD_ICNS]

5. **WASTE**

Name of the dataset: Municipal waste by waste management operations (env_wasmun)
Availability: 2010-2019
Geography: 28 eu countries
Frequency: yearly
Unit of measures: Kilograms per capita [KG_HAB] - Tonne [T] - Thousand tonnes [THS_T]
Use in analysis: Yes
Data description: Municipal waste is mainly produced by households, similar wastes from sources such as commerce, offices and public institutions are included.

The amount of municipal waste generated consists of waste collected by or on behalf of municipal authorities and disposed of through the waste management system. The amount of municipal waste treatment is reported for the treatment operations incineration (with and without energy recovery), recycling, composting and landfilling. Data are available in thousand tonnes and kilograms per person. Wastes from agriculture and from industries are not included. For further detail on the definition please refer to section 3.4.

The Sustainable Development Indicator on municipal waste is expressed in kilograms per capita.