Forecasting Carbon Emissions Across Continents

Andrey Bessalov

ocean

Forecasting Carbon Emissions Across Continents

1. Challenge description
2. Data preprocessing
3. Statistics and visualization
    1. Total emissions by continent
    2. Emissions changes by continent
    3. Total emissions by sector
    4. Emissions changes by sector
    5. GHG types by continent
    6. GHG sectors by continent
    7. GHG types by sector
    8. Analysis of macroeconomic indicators
4. Correlation analysis
    1. GHG and CO2 emissions by countries
    2. GHG and CO2 emissions by sectors
    3. CO2 emissions per capita and GDP
    4. GHG emissions per capita and GDP
    5. CO2 emissions per capita and GDP – part 2
    6. GHG emissions per capita and GDP – part 2
5. Prediction models development
    1. Modeling tasks description
    2. Feature engineering
    3. Modeling results for CO2 emissions
    4. Model predictions of CO2 emissions
    5. Modeling results for GHG emissions
    6. Model predictions of GHG emission

1. Challenge description

The purpose of this challenge is to develop predictive models for future carbon emissions on a continental scale. To execute, participants will need to analyze trends and patterns in historical data to identify key drivers of emissions in each region and apply these findings to indicators by continent. Beyond the data set provided, it is critical to assess the impact of different sectors and economic activities on emission levels.

Exploratory Data Analysis (10 points):

• Perform in-depth exploratory data analysis (EDA) on the dataset, categorizing outcomes and insights based on continents. Thoroughly examine each continent's key statistical measures, patterns, and trends. The dataset spans countries across North America, South America, Europe, Asia, Africa, and Australia. Furthermore, elaborate on the details of your data cleaning and preprocessing approaches, with a specific emphasis on addressing missing values and outliers, and ensuring the uniformity of the data.

CO2 vs GHG Correlation (5 points):

• Explore the correlations between country-specific total CO2 emissions (measured in Mt CO2/yr) and total greenhouse gas (GHG) emissions (measured in Mt GHGeq/yr). Identify any outliers in the dataset and explain their presence if observed.

Sector Contribution Correlation (5 points):

• Explore the correlation between different sectors and the aggregate CO2 emissions (measured in Mt CO2/yr) as well as greenhouse gas (GHG) emissions (measured in Mt GHGeq/yr). Organize your results based on continent rather than individual countries.

Emissions vs GDP Correlation (5 points):

• Examine the correlation between per capita CO2 emissions (measured in t CO2/cap/yr) and GDP, as well as per capita GHG emissions (measured in t GHGeq/cap/yr) and GDP, spanning individual countries and continents. Detect and explain the presence of any outliers in respective countries, describing their occurrence if observed.

Your Own Correlation (5 points):

• Identify one additional correlation within the datasets that you find relevant. Justify your choice by explaining why this particular correlation was selected. Elaborate on why this correlation is a valuable indicator, highlighting its significance in the data context and its potential insights.

Temporal Analysis (10 points):

• Examine the temporal trends in CO2 and GHG emissions across continents. Identify notable patterns or shifts over time, pinpointing periods marked by significant increases or decreases in emissions. Explain these trends, considering policy changes, economic shifts, technological advancements, or natural events that might have influenced the observed patterns.

Prediction Model (30 points):

• For the machine learning component of this challenge, you are asked to create a model that predicts GHG and CO2 emissions by continent for the next three years. This involves thoroughly considering historical emission trends, sectoral contributions, and per capita emission data within each continent.

• Your model should use the grouped data by continent and be trained to identify patterns and correlations in the emissions data. The objective is to forecast future emissions accurately, considering any notable trends or anomalies observed in the historical data, such as real-world factors influencing emission trends.

• You are encouraged to explore a range of machine learning techniques, from traditional statistical models to advanced neural networks, to determine the approach that best captures the complexities of the emissions data. The effectiveness of your model will be assessed on its accuracy, robustness, and ability to generalize across different continents.

Report (30 points):

• Present a thorough report synthesizing your data analysis findings and prediction model in a format suitable for technical and non-technical audiences. Tailor your report for individuals passionate about carbon emissions and climate change who may lack a statistical or data analysis background. Envision communicating your results to an audience unfamiliar with data analytics terminology or techniques. Please consult the "Report Guidelines" section for an in-depth overview and guidance on crafting your report.

2. Data preprocessing

- **Validation of input data**
  - We have 4 input files for fossils CO2 emissions and 4 for GHG emissions.
  - The first step was to make sure that the values from the 2 files with totals coincide with the summed values from the files GHG_by_sector_and_country.csv
  - We checked that the values from the fossil_CO2 file match the values from the GHG file with a filter for substance=CO2.
- **Data preprocessing steps**
  - Convert data from wide to long format.
  - Replace the nulls with the average values of the neighbors, i.e. (bfill + ffill) /2.
  - For some small countries there are no complete indicators for GDP and population. We will not take these countries into account since their contribution is small.
- **Continents description**
  - We merged the continents taken from Wikipedia: https://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_continent_(data_file)
  - For some countries there is uncertainty in defining the continent, in this case, we determined it expertly.
  - Definitions: AF – Africa, AS – Asia, EU – Europe, NA – North America, OS – Australia and Oceania, SA – South America.
- **Sectors description (retrieved from the EDGAR paper: https://edgar.jrc.ec.europa.eu/report_2023?vis=co2tot#emissions_table)**
  - Power industry includes power and heat generation plants (public and auto-producers).
  - Industrial combustion and processes includes combustion for industrial manufacturing and industrial process emissions (e.g. non-metallic minerals, non-ferrous metals, solvents and other product use, chemicals, etc.).
  - Transport includes road transport, rail transport, domestic aviation, domestic shipping and inland waterway transport for each country.
  - International shipping and aviation also belong to this sector and are presented separately in the country factsheets due to their international nature.
  - Buildings includes small-scale non-industrial stationary combustion.
  - Agriculture includes agriculture livestock (enteric fermentation, manure management), agriculture soils (fertilisers, lime application, rice cultivation, direct soil emissions, indirect N2O emissions from agriculture), field burning of agricultural residues.
  - Waste includes solid waste disposed on land, solid waste composted and hazardous solid waste processing/storage, waste water handling, waste incineration.
  - Fuel exploitation: fuel extraction, transformation and refineries activities, including venting and flaring.

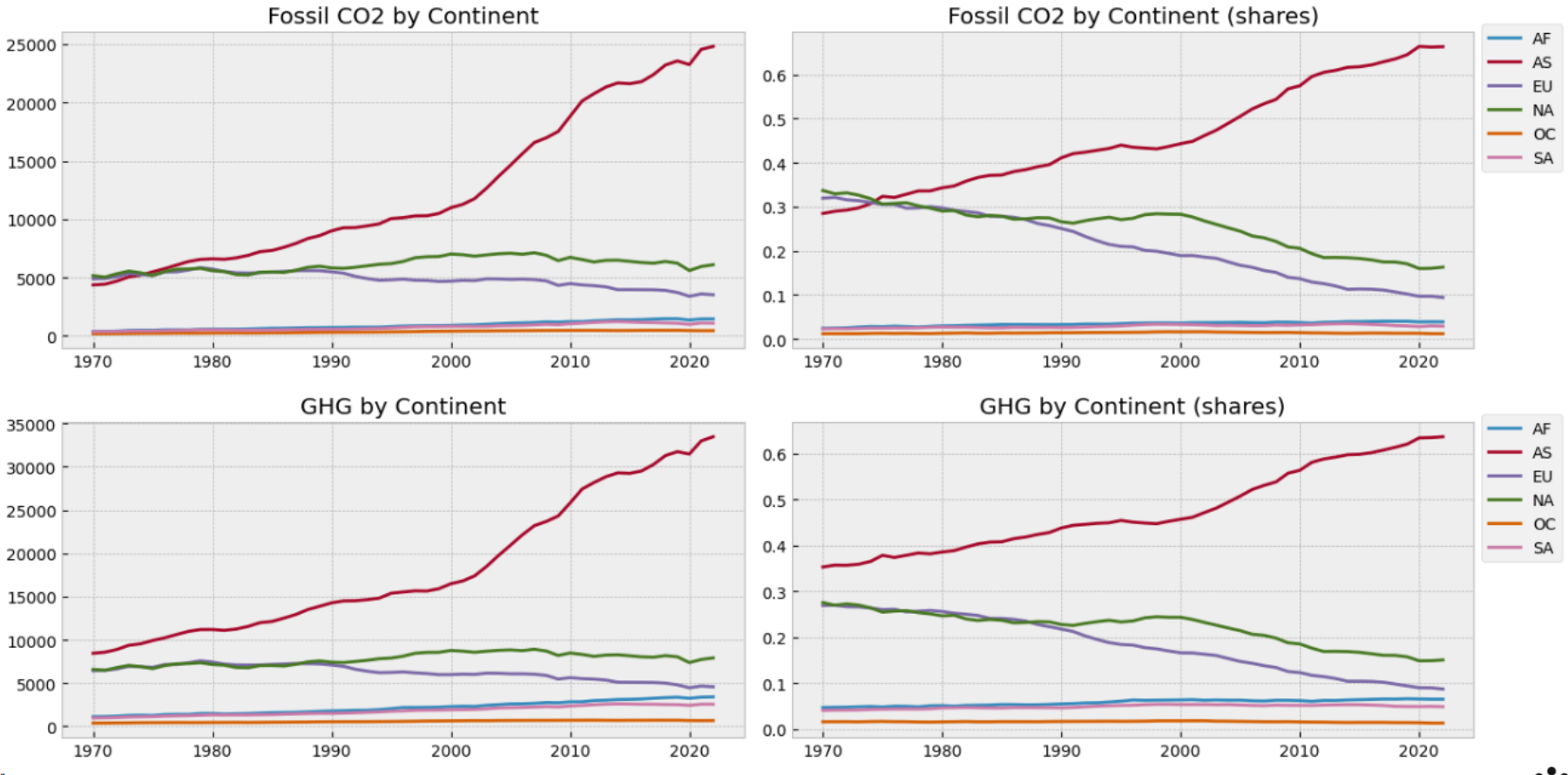
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/1_preprocessing.ipynb

ocean

3.1. Analysis of total emissions by continent

The graphs below show the absolute values of continental CO2 and GHG emissions expressed in Mt CO2/yr on the left, and the shares of these values from the total contribution on the right.

- We see a very high correlation between fossil CO2 and GHG values.
- The most polluting continent is Asia. In recent years, the values of emissions reach 33 Gt GHG per year and 25 Gt CO2 per year, which is approximately 2/3 of the total emissions of the entire earth.
- Australia is the least polluting continent.
- In Asia there is an increase in emissions, while in Europe and North America there has been a decline in recent years.



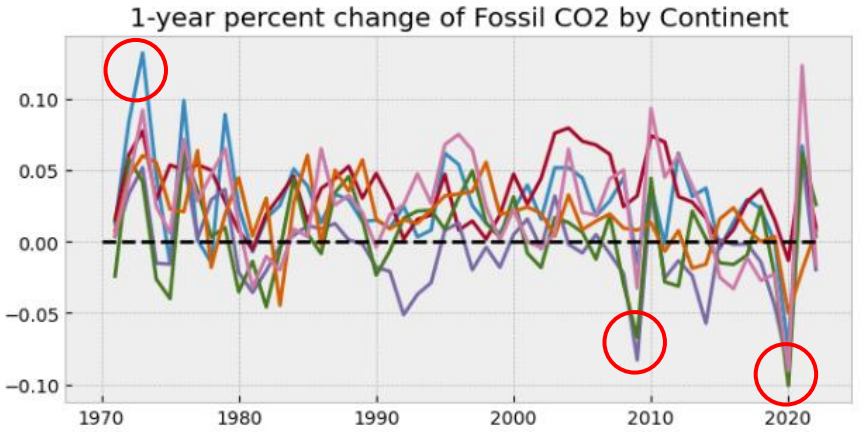
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

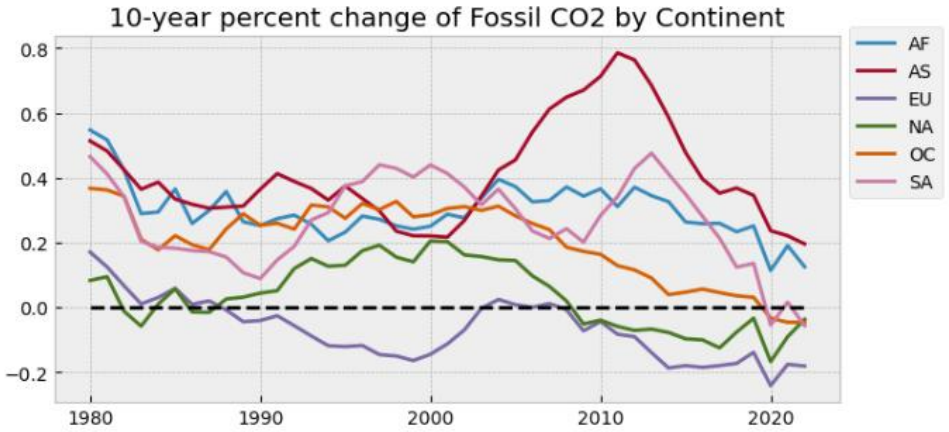
ocean

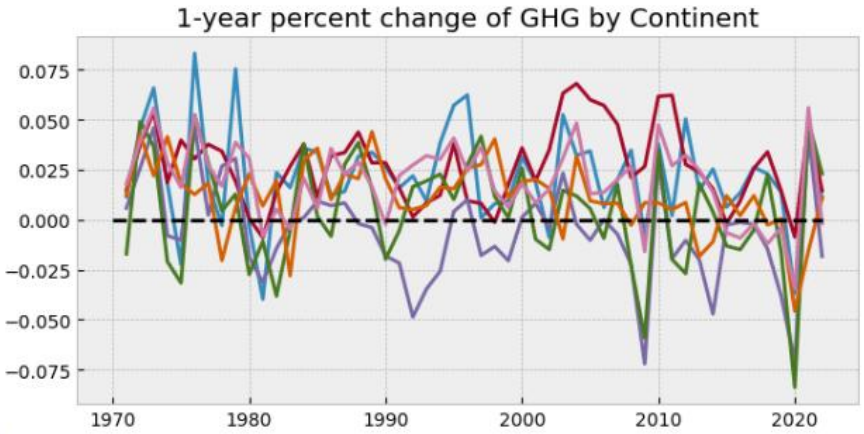
3.2. Analysis of emissions changes by continent

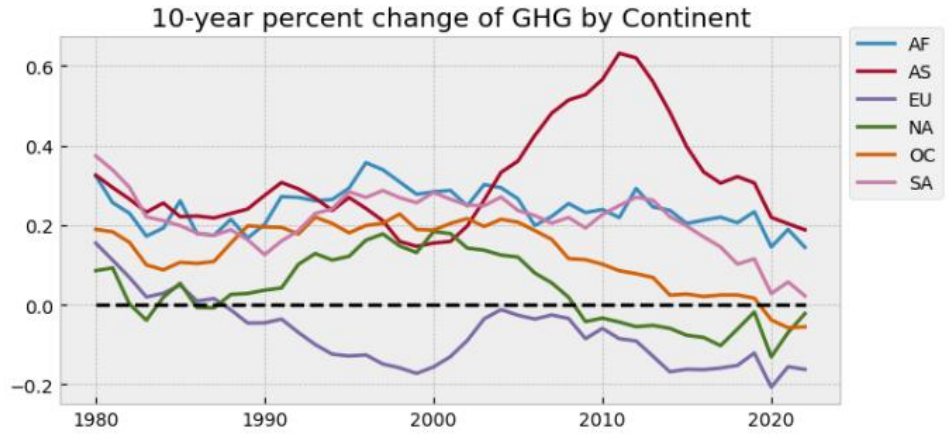
The graphs below show the percentage changes in continental CO2 and GHG emissions over a year on the left, and over 10 years period on the right.

- We see that in 2020, due to COVID-19, there was a sharp drop in emissions, with CO2 emissions falling more than the GHG.
- In 2008, there was a decline in emissions in North America and Europe, likely due to the financial crisis in that time.
- We see a series of sharp increases in Africa in the 70s, probably related to the specific of input data.
- From the 10 year changes plots, it can be noted that Europe and North America have been reducing their emissions over the past 10-15 years. In Asia, emissions are increasing, while the growth rate is falling (it was at maximum levels from 2000 to 2012).









* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

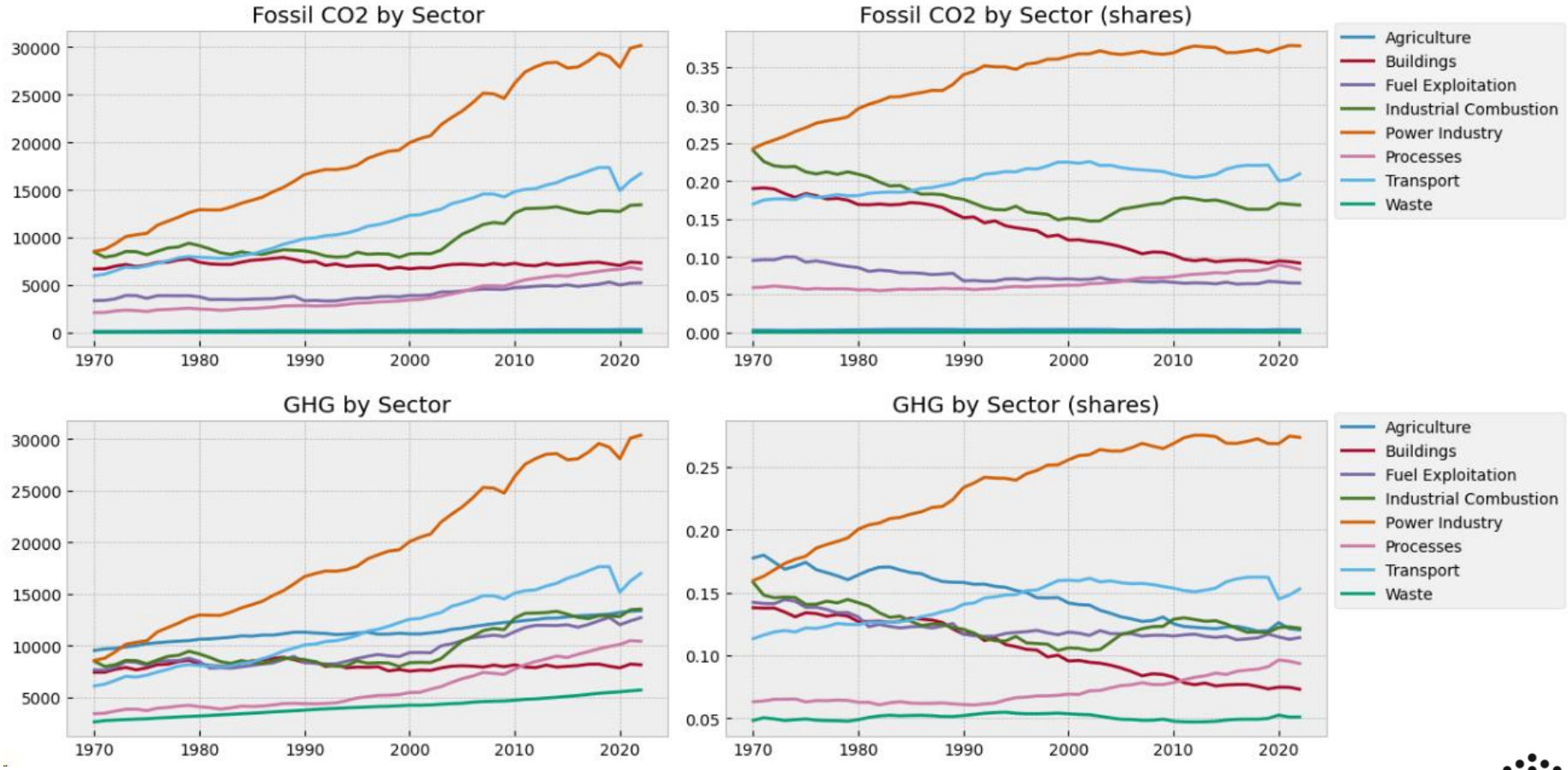
ocean

3.3. Analysis of total emissions by sector

The graphs below show the absolute values of sectoral CO2 and GHG emissions expressed in Mt CO2/yr on the left, and the shares of these values from the total contribution on the right.

• The most polluting sector is Power Industry. In recent years, the value of CO2 emissions reaches 30 Gt per year, which is approximately 30% of total GHG emissions, and 40% of total CO2 emissions.

• The least polluting sectors in terms of GHG emissions is Waste.



* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

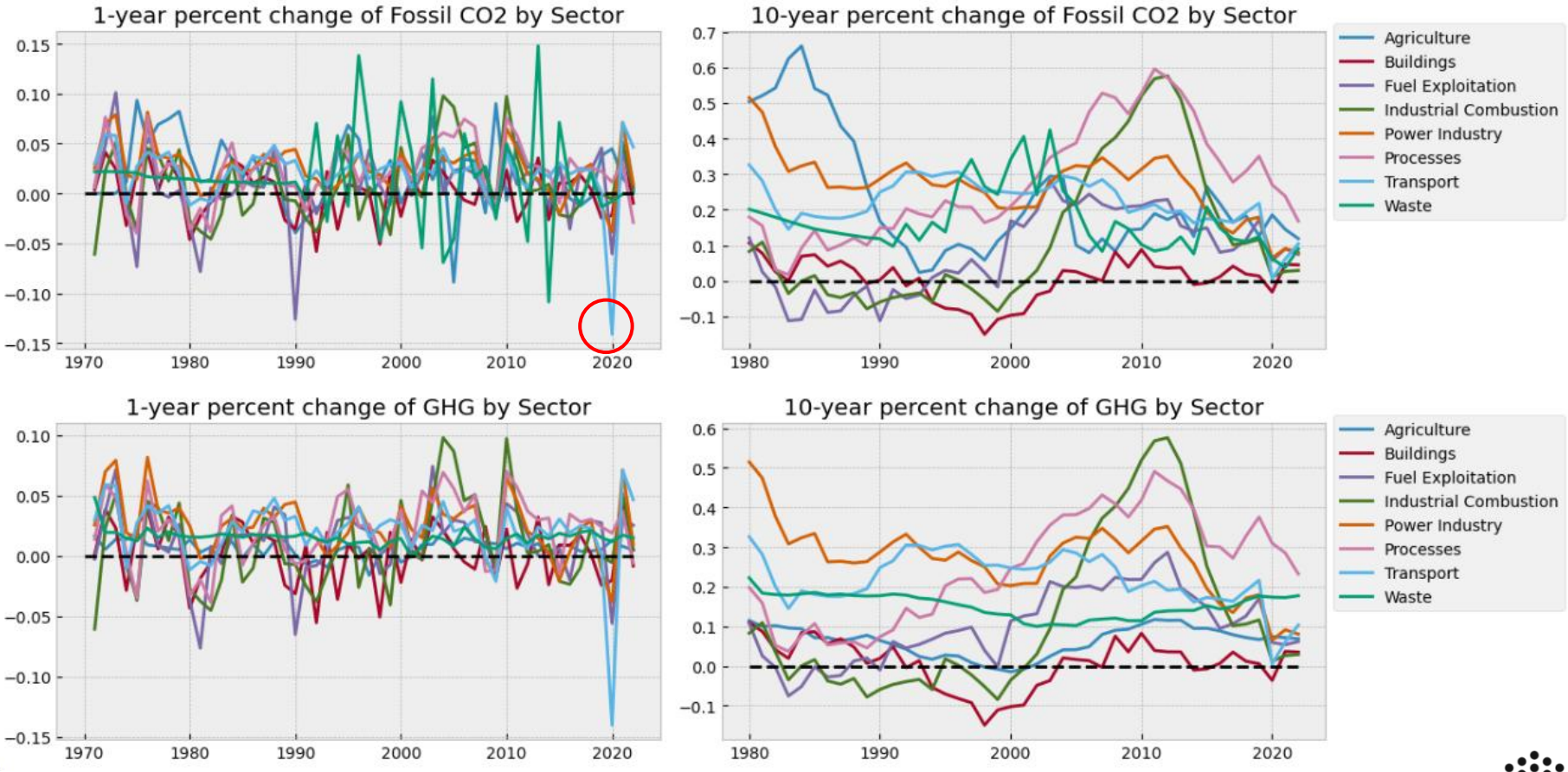
3.4. Analysis of emissions changes by sector

The graphs below show the percentage changes in sectoral CO2 and GHG emissions over a year on the left, and over 10 years period on the right.

• We see that in 2020, due to COVID-19, there was a sharp drop in emissions in the Transport sector, which is explained by the fact that people moved less and stayed at home during this period.

• Large fluctuations in the Waste sector are associated with low CO2 values for this category.

• Based on 10 year changes, we see that all sectors are growing in emissions, the greatest growth is in Processes, but the intensity of this growth is decreasing.



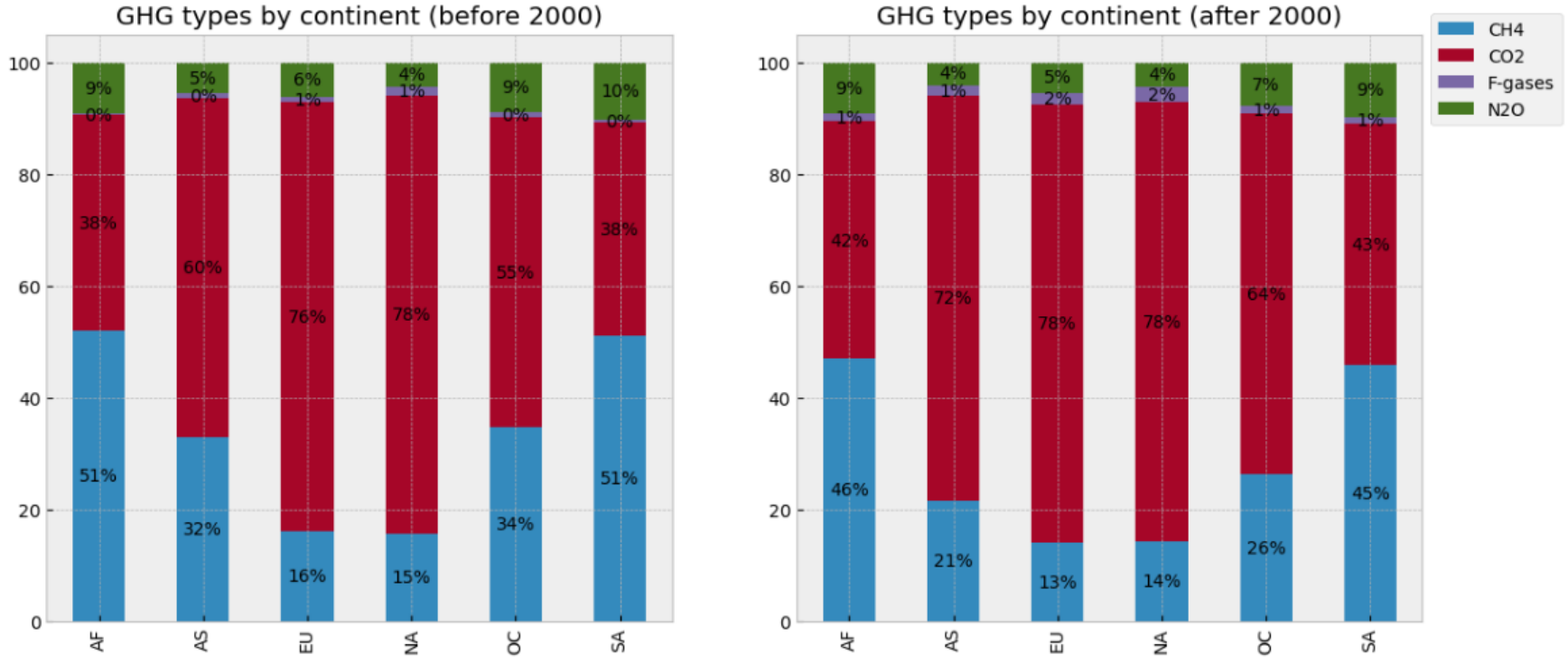
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

3.5. Analysis of GHG types by continent

The graphs below show the shares of GHG types by continent.

- According to the degree of CO2 contribution to total emissions, countries can be divided into 3 groups:
  - Europe and North America;
  - Asia and Australia;
  - Africa and South America;
- We see that since 2000 the distribution picture has changed slightly, with some highlighted changes:
  - The share of CO2 increased for every continent except Europe and North America.
  - Asia becomes more similar to Europe and North America.



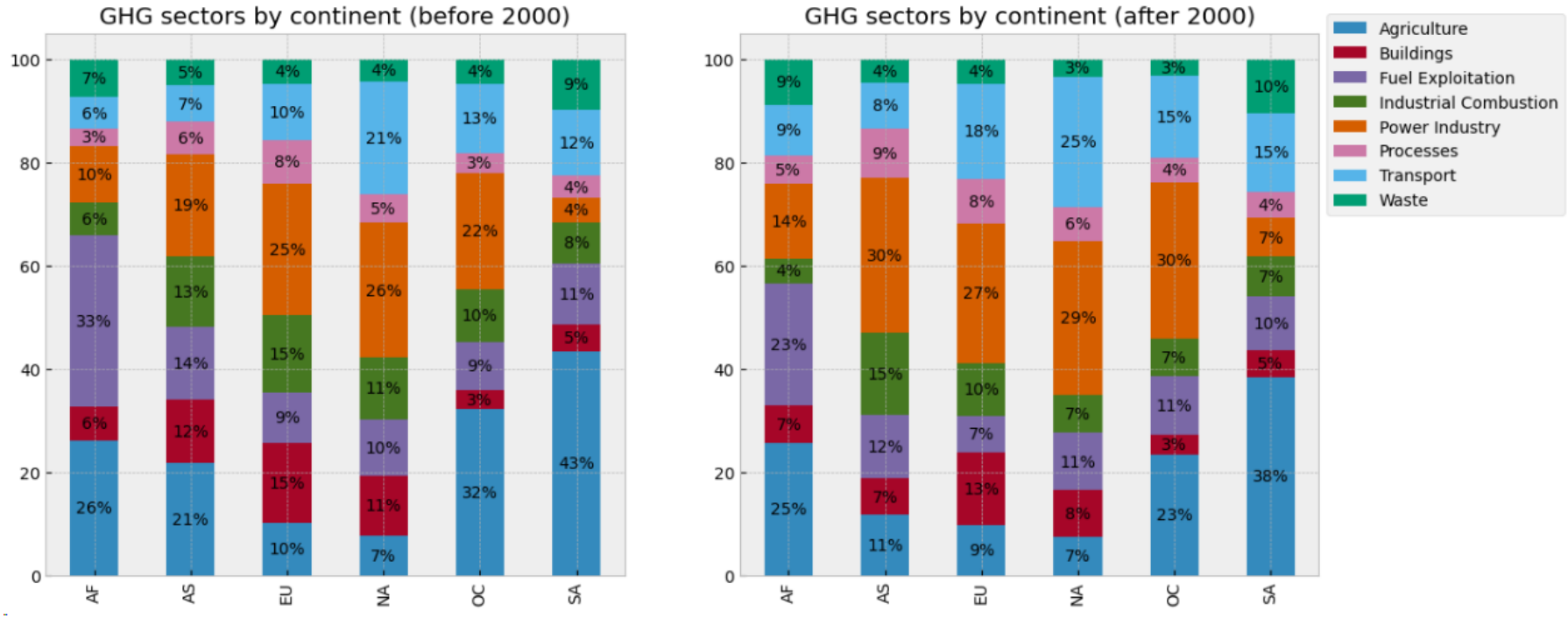
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

3.6. Analysis of GHG sectors by continent

The graphs below show the shares of GHG sectors by continent.

- In terms of the sector's contribution to the total emissions in all continents except Africa and South America, the largest contribution comes from Power Industry. In Africa Agriculture and Fuel Exploitation are two main sectors, in South America - Agriculture is the most pollutant sector.
- We see that since 2000 the distribution picture has changed slightly, but there are still some changes:
  - In Africa has dropped Fuel Exploitation.
  - In Asia has grown Power Industry.
  - In Europe Transport has grown.



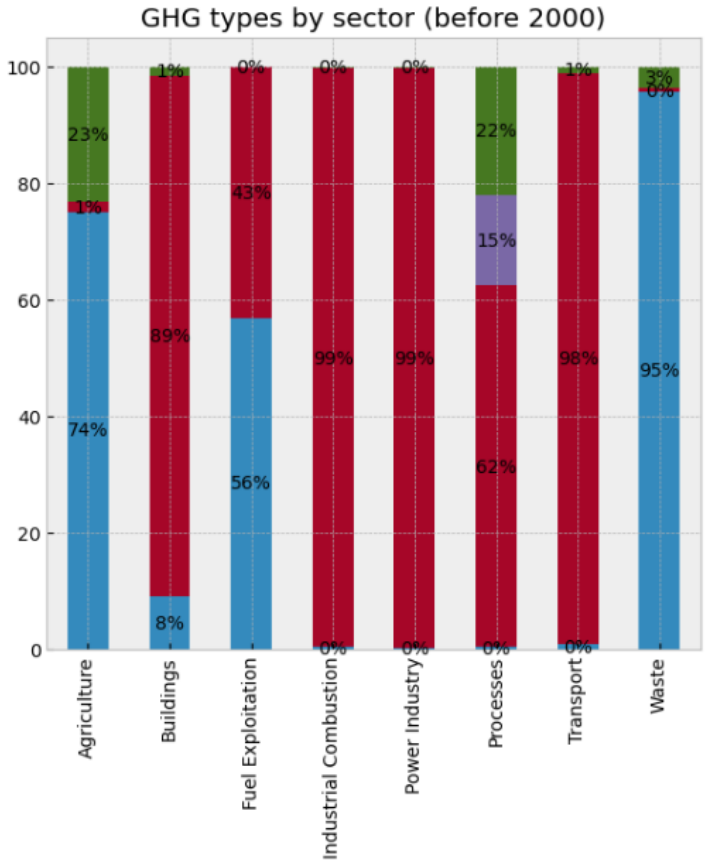
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

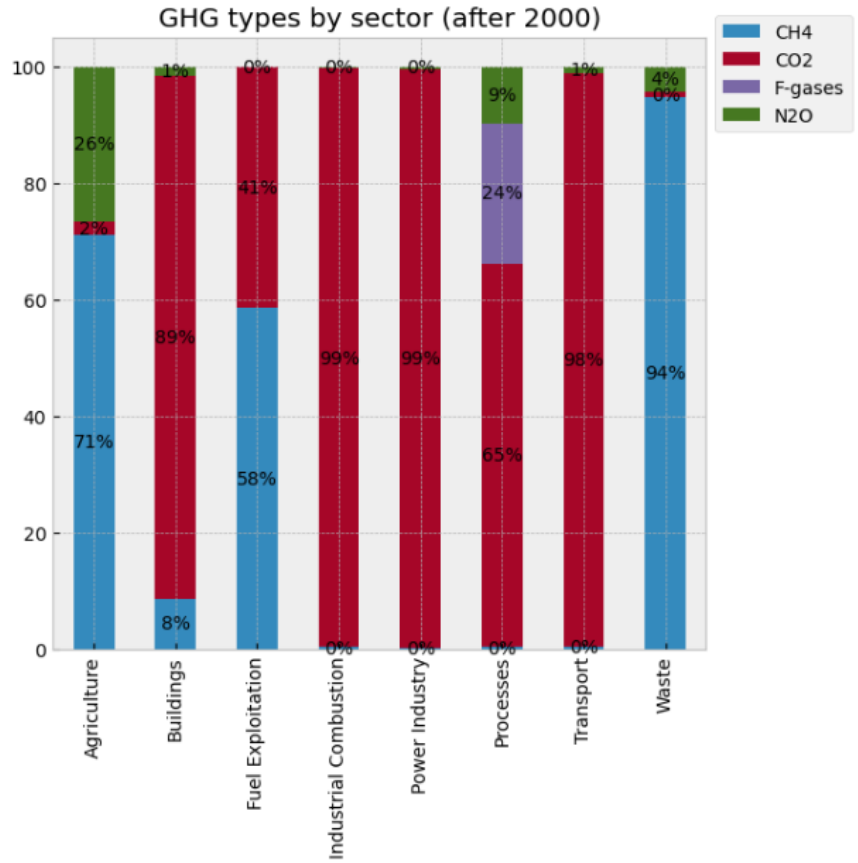
ocean

3.7. Analysis of GHG types by sector

The graphs below show the shares of GHG types by sector. We can break down sectors according to the types of gases that are released there.

- Almost only CO2 is released from Industrial Combustion, Power Industry and Transport sectors.
- In Waste and Agriculture sectors there is usually CH4 (methane).
- In Buildings usually CO2.
- In Fuel Exploitation CO2 and methane (methane a little more).
- In the Process, as a rule, CO2, but also a large proportion of F-gases.





* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

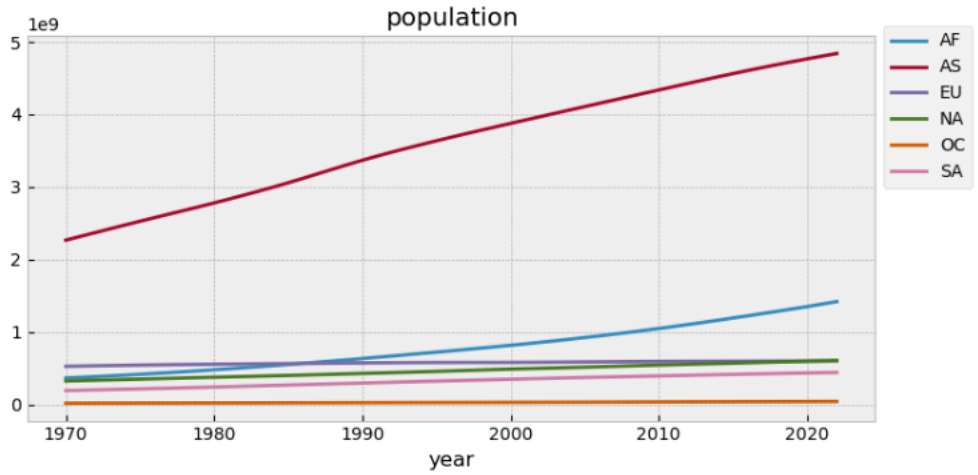
ocean

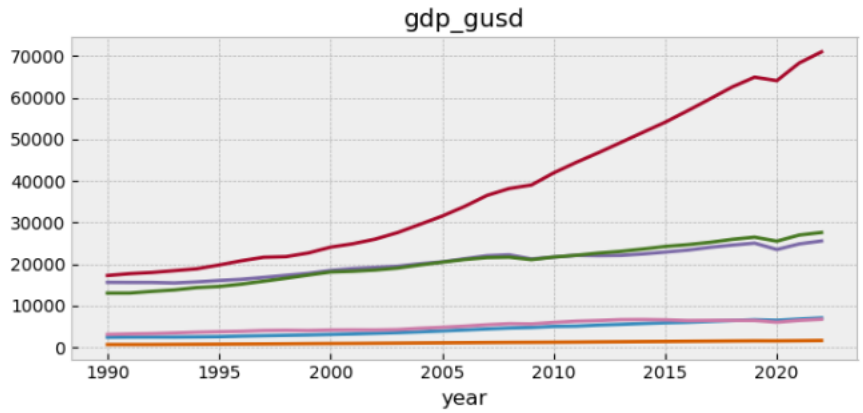
3.8. Analysis of macroeconomic indicators

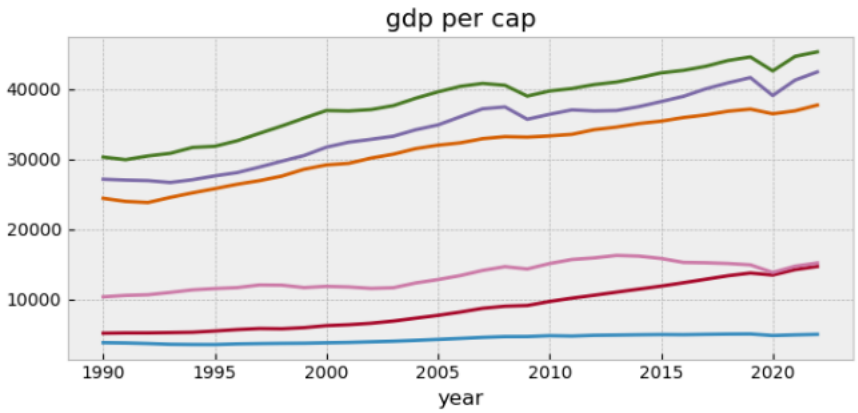
The graphs below the main macroeconomic indicators that we will use to build a CO2 and GHG forecasting models. We derived these indicators from the input files by counting them at the country level and then aggregating to the continent.

• We see that about 5 billion people live in Asia, about 1.5 billion in Africa, and less than a billion on other continents.

• North America has the highest GDP per capita, Africa the lowest. Three groups of continents can be distinguished according to this indicator, where the values are approximately in the same range. Rich (North America, Europe, Australia), average (Asia, South America), poor (Africa).







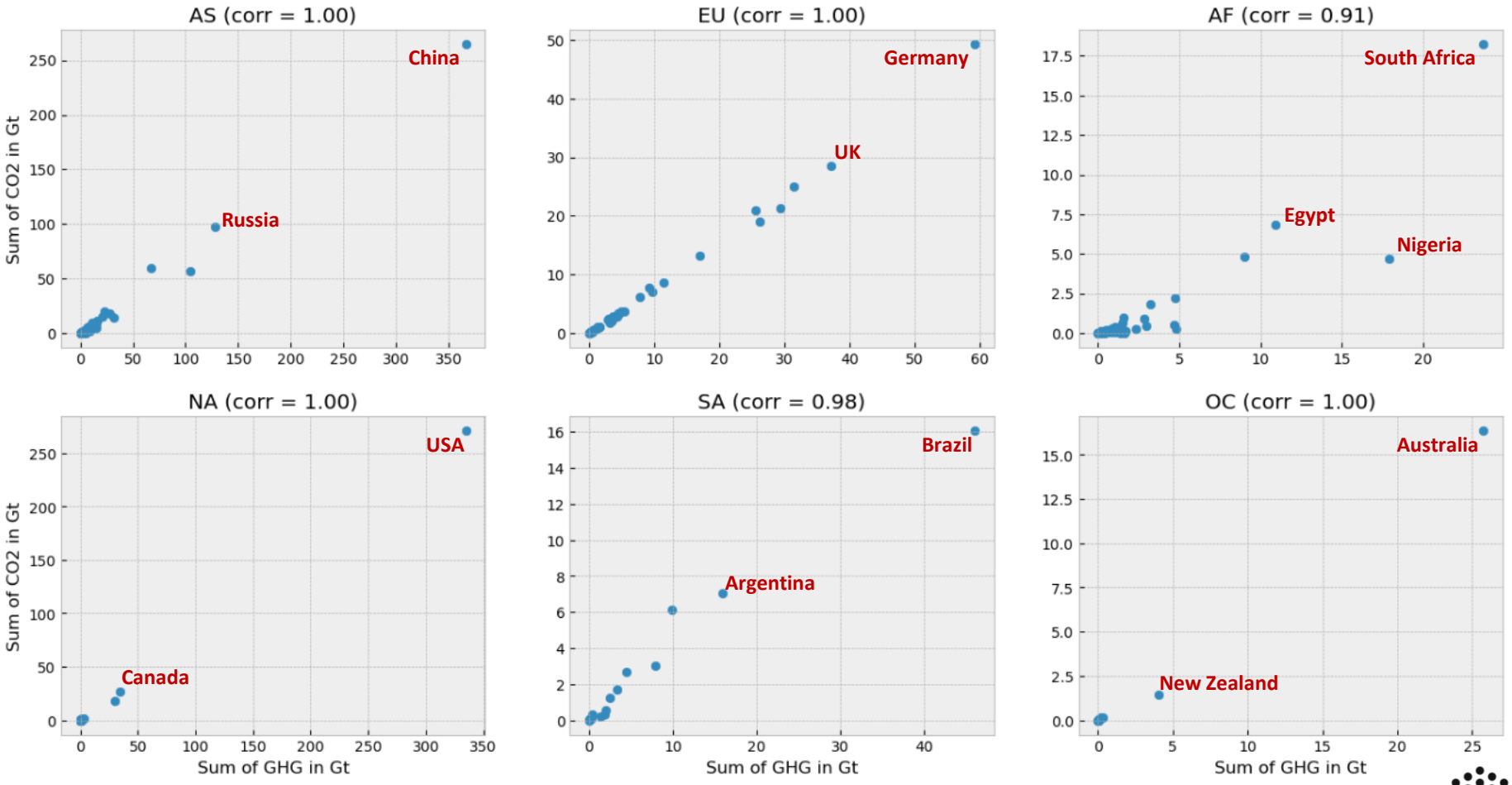
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

4.1. Correlation analysis of GHG and CO2 emissions by countries

The graphs below show the dependences of total CO2 and GHG emissions summed up for all time by country, expressed in Gt.

- We see that across all continents the dependence between these indicators is very high (Pearson's correlation coefficient is close to 1).
- The lowest rate is observed in Africa.



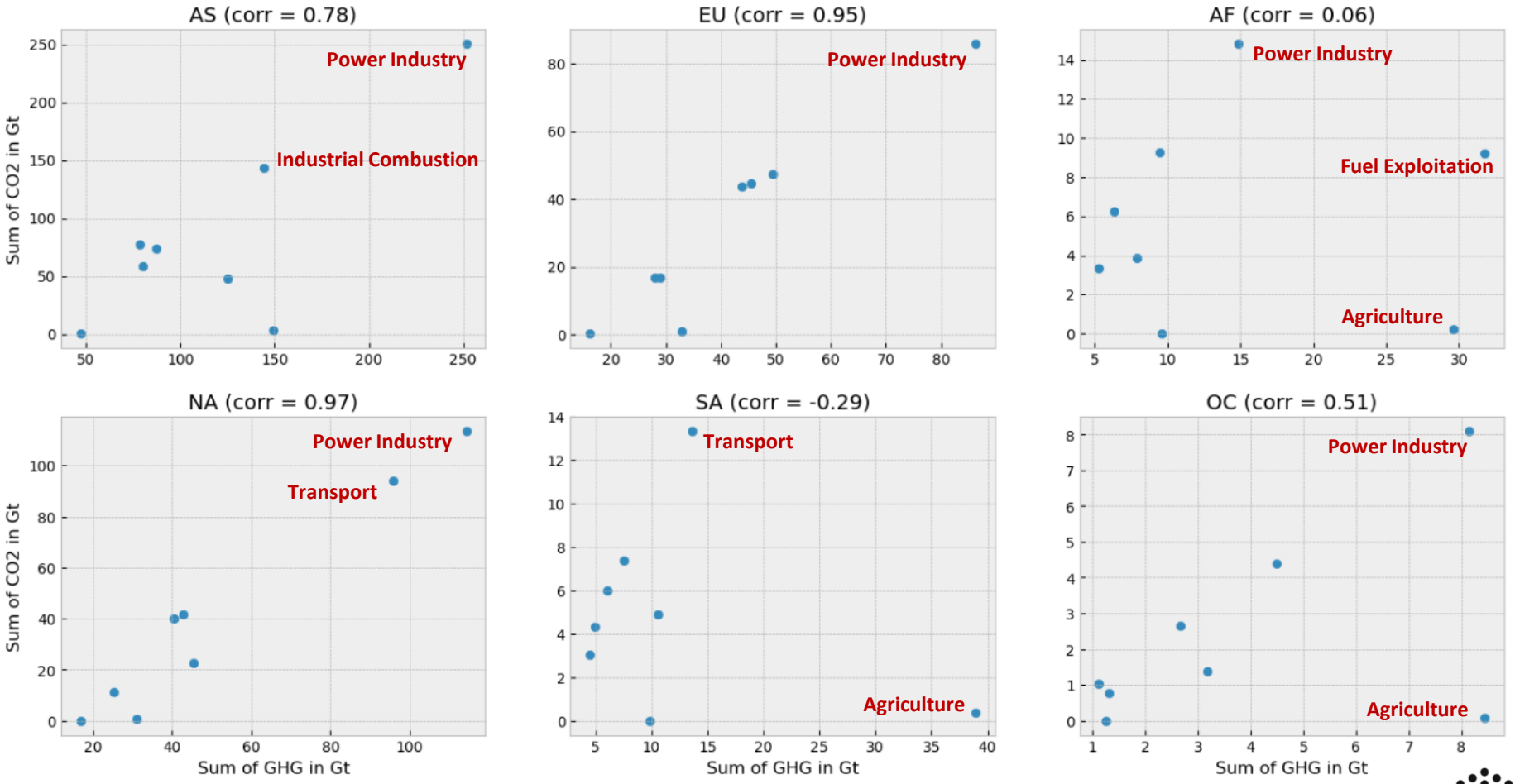
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

4.2. Correlation analysis of GHG and CO2 emissions by sectors

The graphs below show the dependences of total CO2 and GHG emissions summed up for all time by sector, expressed in Gt.

- We see that in Europe and North America the dependence between these indicators is very high (Pearson's correlation coefficient is close to 1).
- For Asia and Australia the correlation is also high, but not close to 1.
- For Africa and South America the correlation is low due to the presence of some outliers in the data.

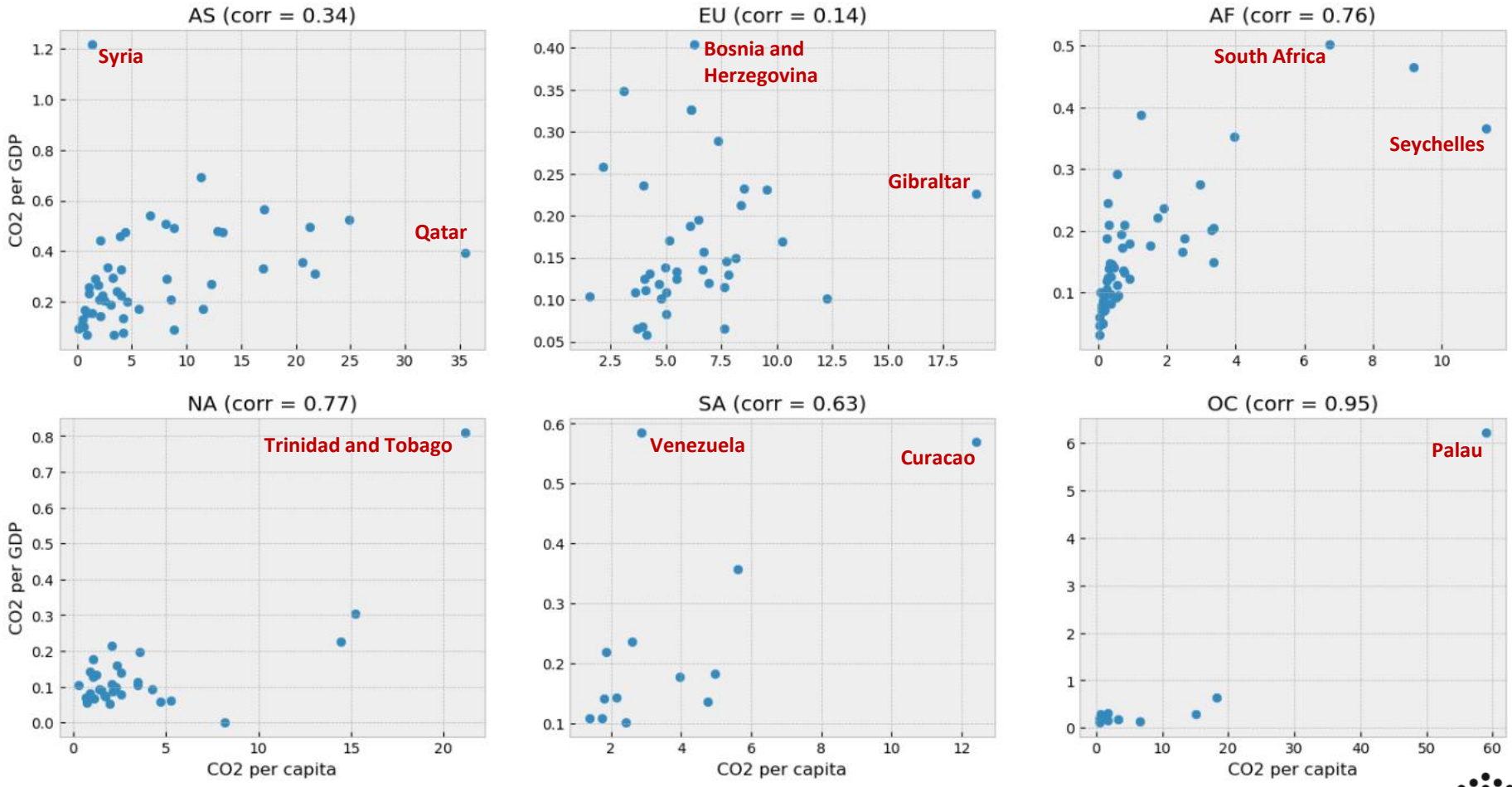


* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

4.3. Correlation analysis of CO2 emissions per capita and GDP

The graphs below show the dependences of CO2 emissions per capita and GDP calculated for 2022. For some continents we see a strong dependence between these indicators, for others the Pearson correlation coefficient is low due to the presence of outliers in the data.

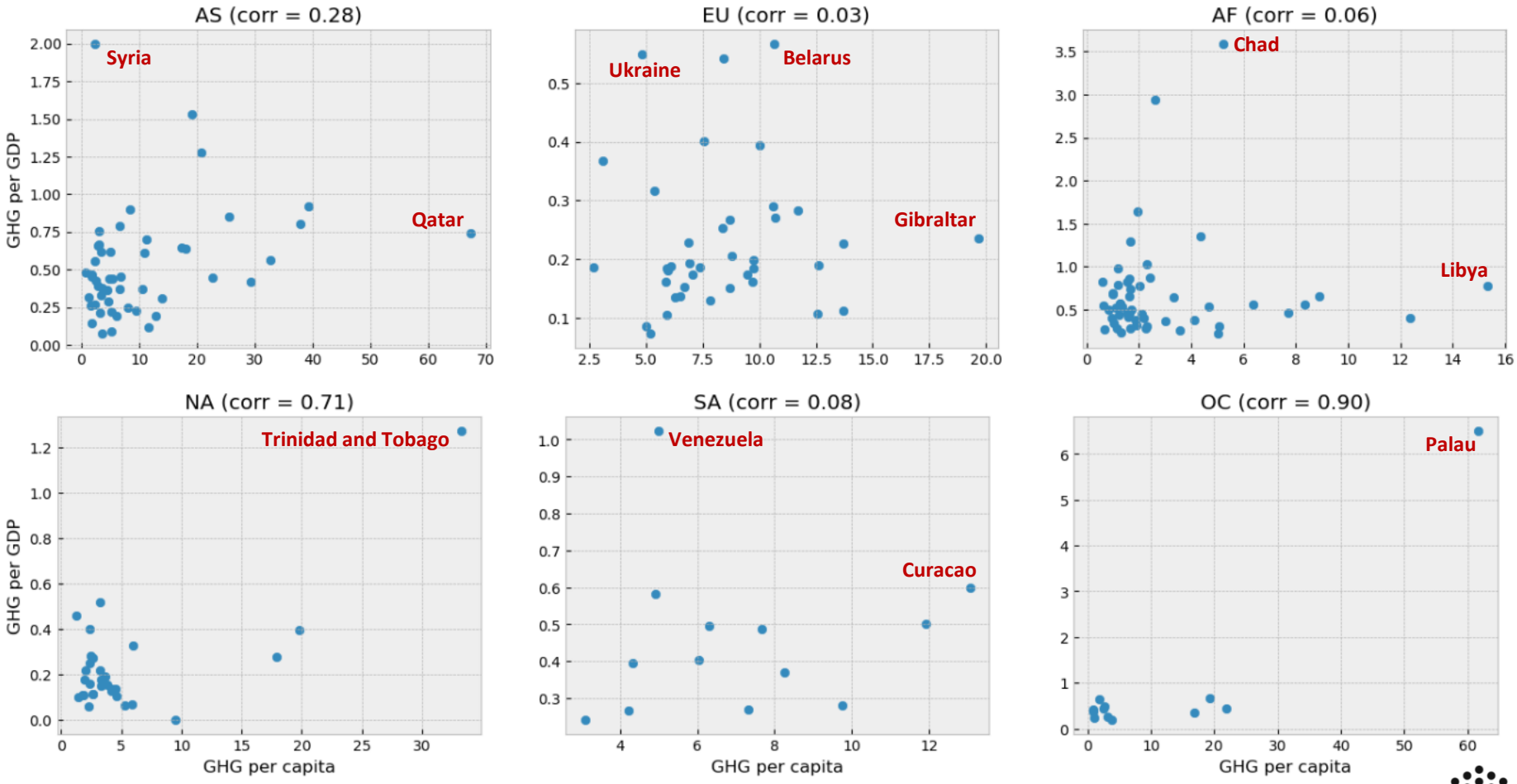


* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

4.4. Correlation analysis of GHG emissions per capita and GDP

The graphs below show the dependences of GHG emissions per capita and GDP calculated for 2022. For some continents we see a strong dependence between these indicators, for others the Pearson correlation coefficient is low due to the presence of outliers in the data.

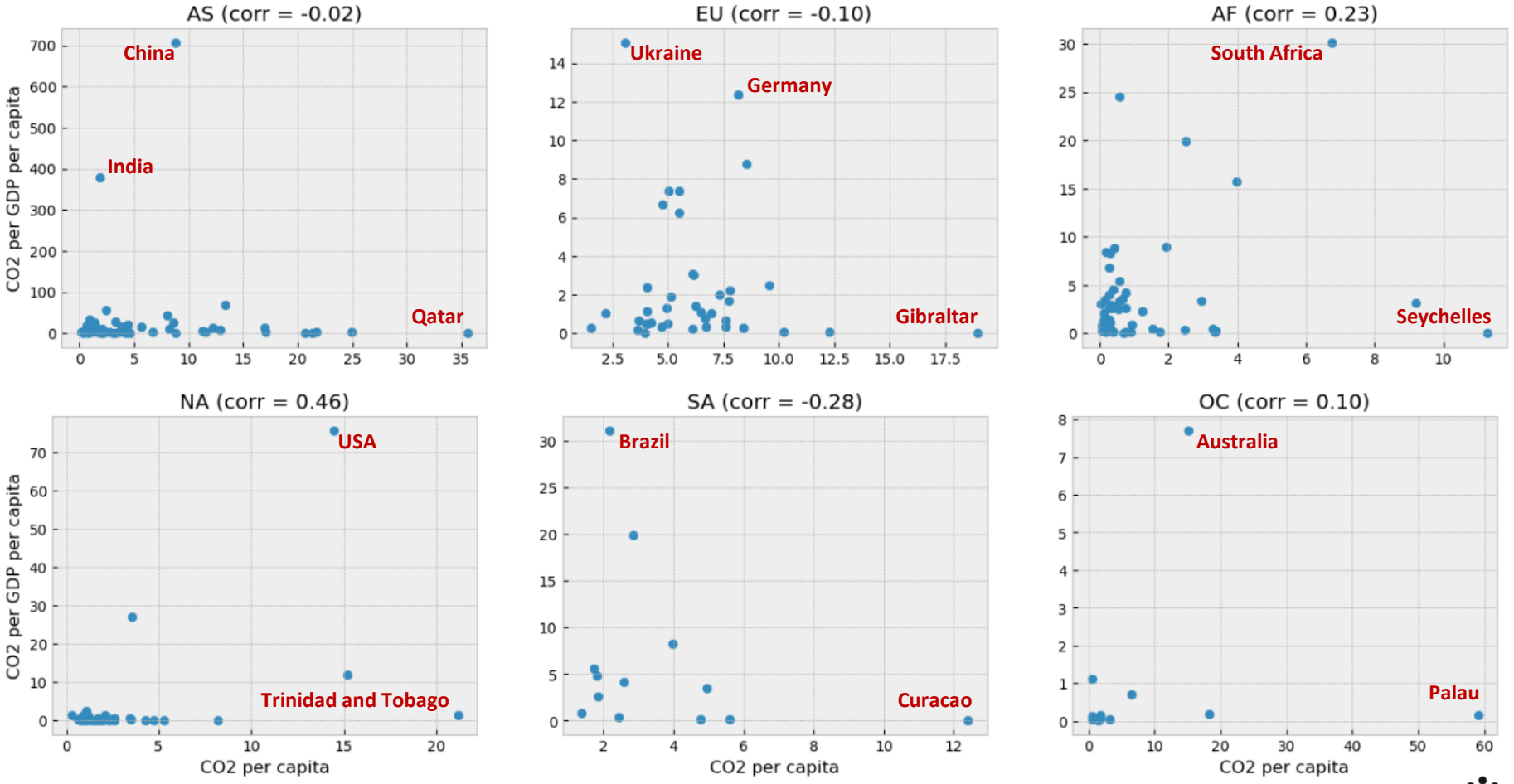


* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

4.5. Correlation analysis of CO2 emissions per capita and GDP – part 2

The graphs below show the dependences of CO2 emissions per capita and GDP per capita calculated for 2022. For some continents we see a strong dependence between these indicators, for others the Pearson correlation coefficient is low due to the presence of outliers in the data.

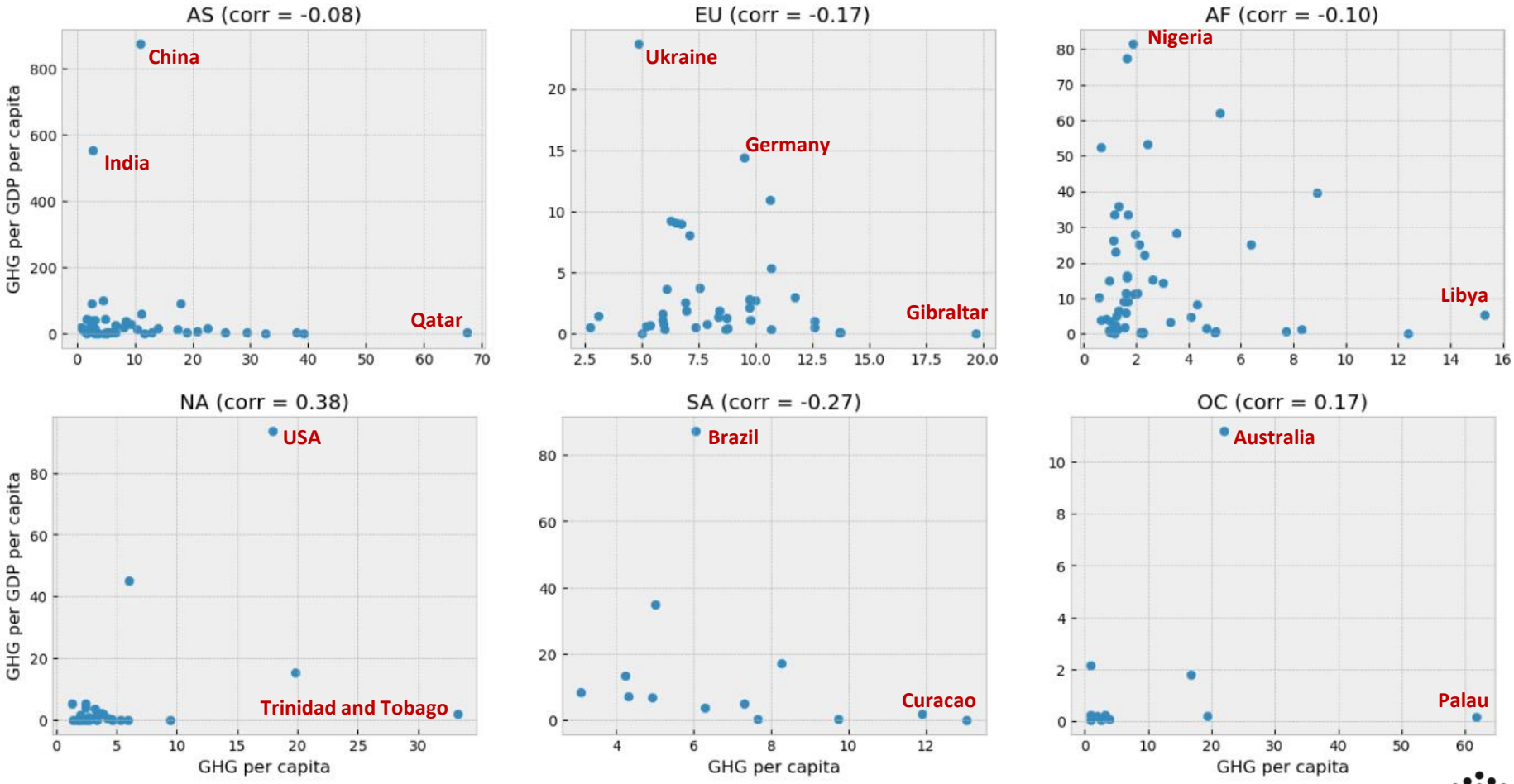


* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

4.6. Correlation analysis of GHG emissions per capita and GDP – part 2

The graphs below show the dependences of GHG emissions per capita and GDP per capita calculated for 2022. For some continents we see a strong dependence between these indicators, for others the Pearson correlation coefficient is low due to the presence of outliers in the data.



* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/2_statistics.ipynb

ocean

5.1. Modeling tasks description

In this section we will develop models for predicting the overall CO2 and GHG emissions by every continent for the next 3 years.

Let's set the following goals:
• To build different models find ones giving the best results on the validation sets.
• To get predictions using the best models for every continent.

Validation:
• As the validation set we took 3 folds of 3 years old. Each time the model forecasts 3 years ahead using all past data.
• As a quality metrics we took MAPE – mean absolute percentage error.

Models tested:
Since we don’t have a lot of data but only yearly aggregations we used models not sensitive to overfitting.
• Random Forest Regressor.
• Elastic Net Regressor with best parameters searching.

Target transformations:
At first we replaced anomaly target values in years 2008 and 2020 by their corresponding neighbors mean values.
Then we tested different target transformations before modeling and chose ones giving the best MAPE on the validation sets:
• Without transformation.
• Differences(1).
• Differences(1) + LocalStandardScaler.

* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/3_training.ipynb

ocean

5.2. Feature engineering

The table below presents a list of fields generated separately for Fossil CO2 and GHG emissions by every continent used to predict their future values for the next 3 years.

• We use lagged variables and rolling means of that lags.

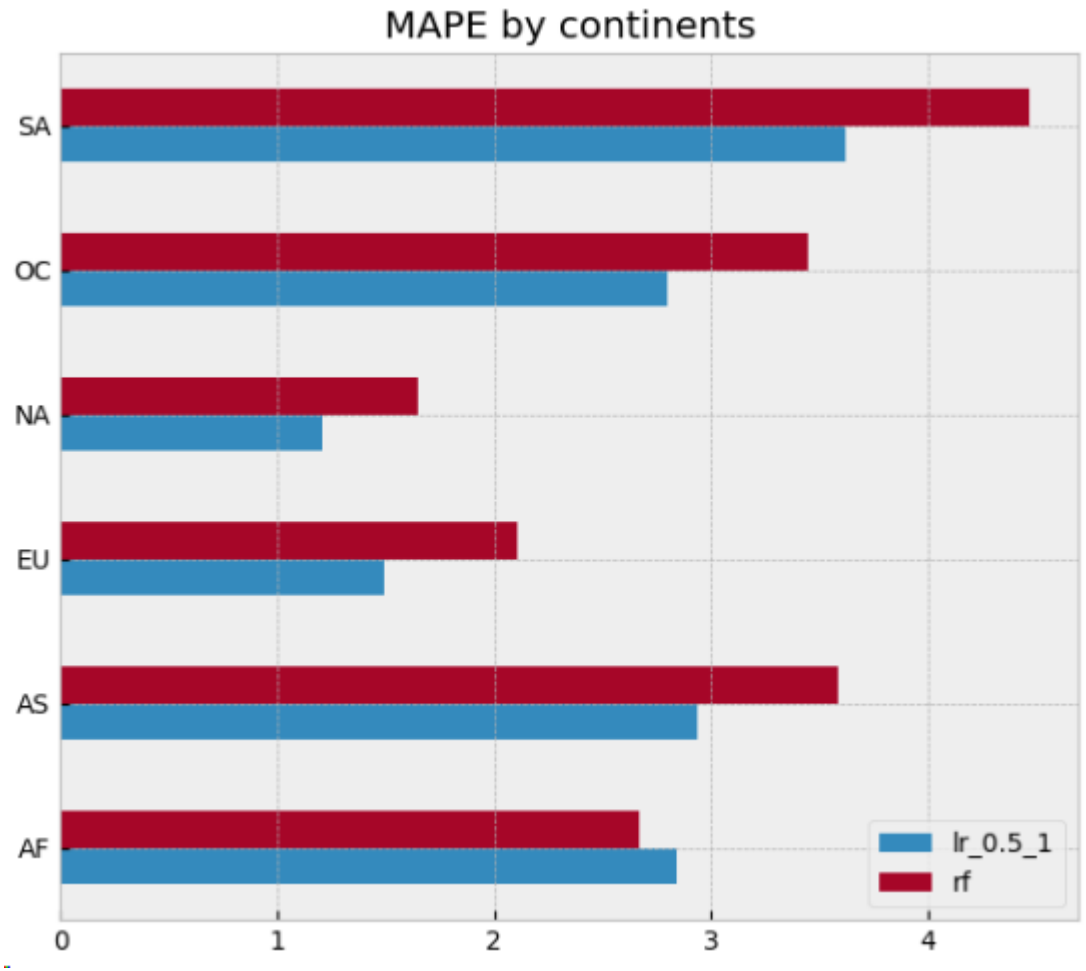
• For Sectoral contributions and Macroeconomic features we don't know their future values so we generate features based on their 3 years-back values.

Area	Feature	Description
Lags of the target	lag1	First lag of the target variable.
	lag2	Second lag of the target variable.
	lag3	Third lag of the target variable.
	RM_lag1_WS3	Rolling mean of first 3 lags to get smoothed value.
Sectoral contributions	y_Agriculture	The share of Agriculture sector from all emissions.
	y_Buildings	The share of Buildings sector from all emissions.
	y_Fuel Exploitation	The share of Fuel Exploitation sector from all emissions.
	y_Industrial Combustion	The share of Industrial Combustion sector from all emissions.
	y_Power Industry	The share of Power Industry sector from all emissions.
	y_Processes	The share of Processes sector from all emissions.
	y_Transport	The share of Transport sector from all emissions.
	y_Waste	The share of Waste sector from all emissions.
Macroeconomic features	population	Continent's share of total population.
	gdp_gusd	Continent's share of total GDP.
	gdp_cap	Continent's share of total GDP per capital.
	val_pop	Continent's proportion by the emissions / population.
	val_gdp_kusd	Continent's proportion by the emissions / GDP.
	val_gdp_cap	Continent's proportion by the emissions / GDP per capital.

* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/3_training.ipynb

5.3. Modeling results for CO2 emissions

• We trained Random Forest Regressor and Elastic Net models with alpha and l1_ratio parameters searching for predicting future values of CO2 emissions.
• We tested different target conversions and found that using first target differences works better.
• As we see on the picture below that for most of the continents Linear Regression (lr_0.5_1) have better results than Random Forest (rf). Overall **MAPE ~ 2.5%** that is a very good result.
• On the right table below we shown the coefficients of the best linear model.



Linear model coefficients:

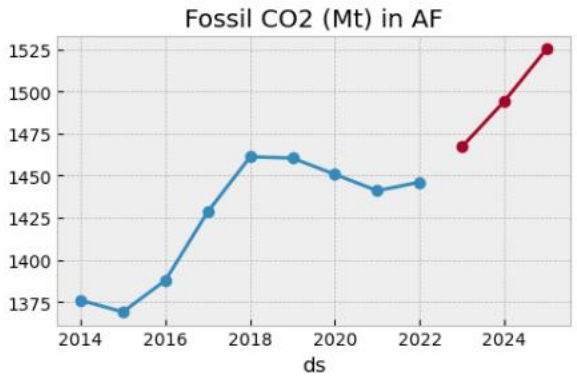
y_Buildings	-105.95
val_gdp_cap	212.87
lag1	0.62
lag2	-0.20
lag3	0.32
RM_lag1_WS3	-0.09

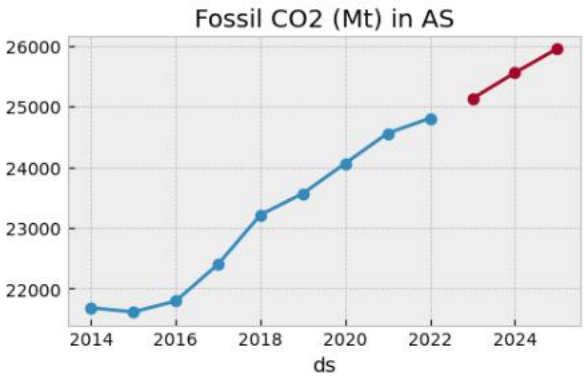
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/3_training.ipynb

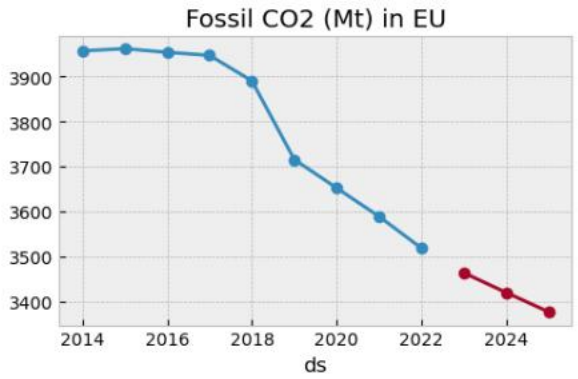
ocean

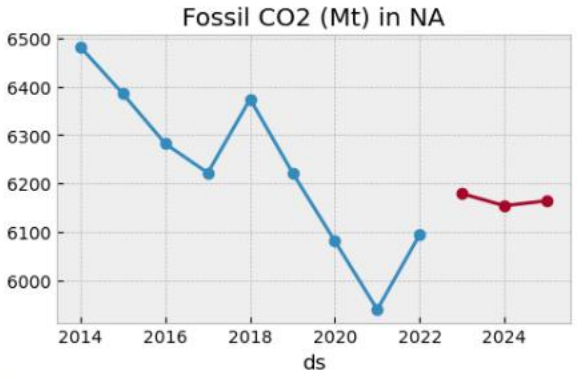
5.4. Model predictions of CO2 emissions

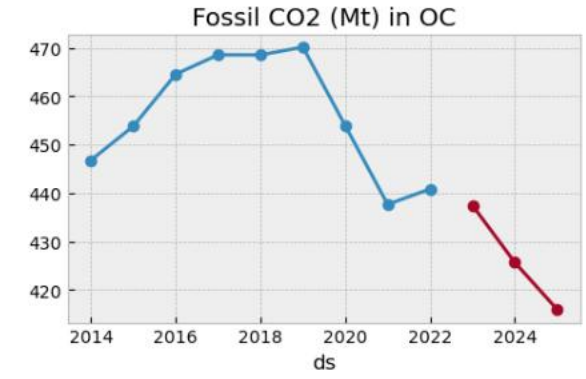
The graphs below show the 3 years of forecasts made by the best model – linear regression Elastic Net.

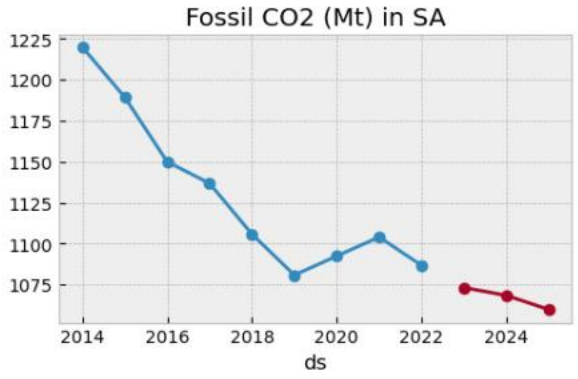










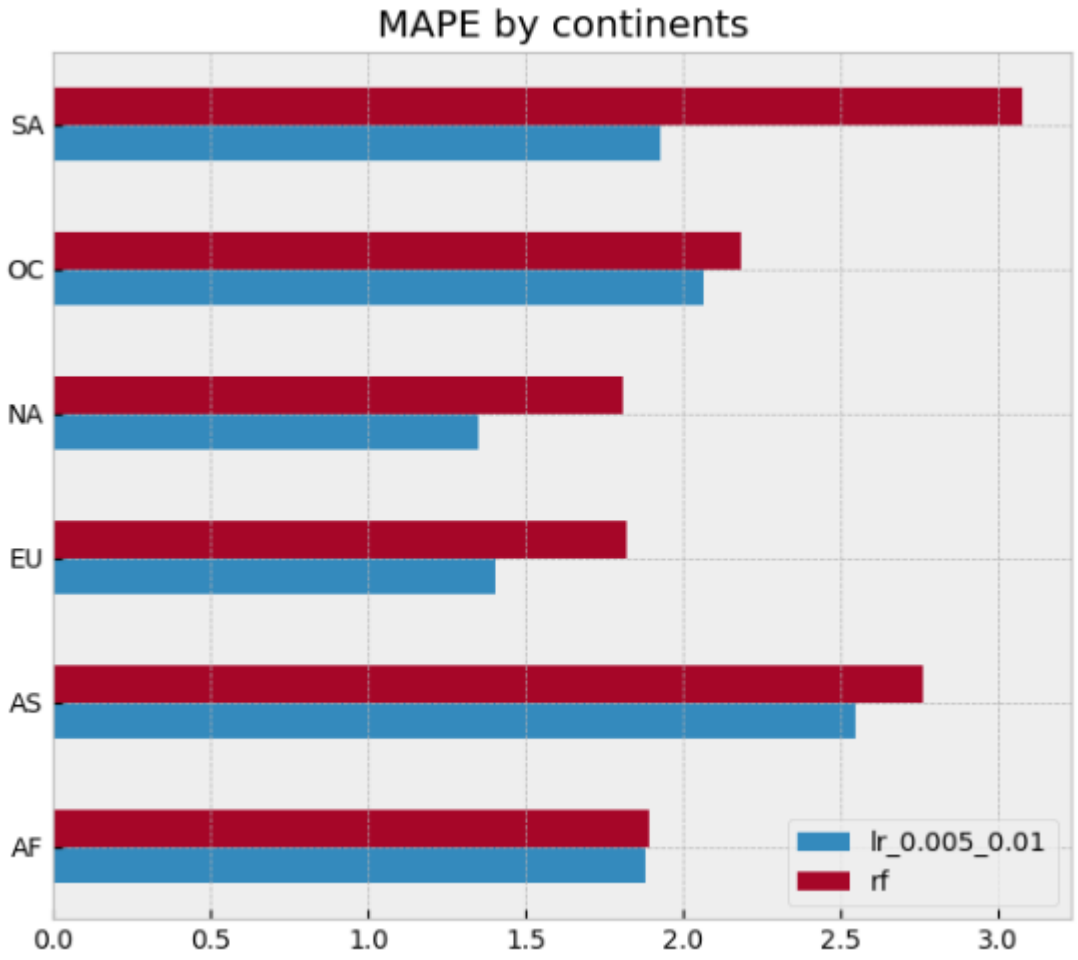


* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/3_training.ipynb

ocean

5.5. Modeling results for GHG emissions

• We trained Random Forest Regressor and Elastic Net models with alpha and l1_ratio parameters searching for predicting future values of CO2 emissions.
• We tested different target conversions and found that using first target differences works better.
• As we see on the picture below that for most of the continents Linear Regression (lr_0.005_0.01) have better results than Random Forest (rf). Overall **MAPE ~ 1.9%** that is a very good result.
• On the right table below we shown the coefficients of the best linear model.



Linear model coefficients:

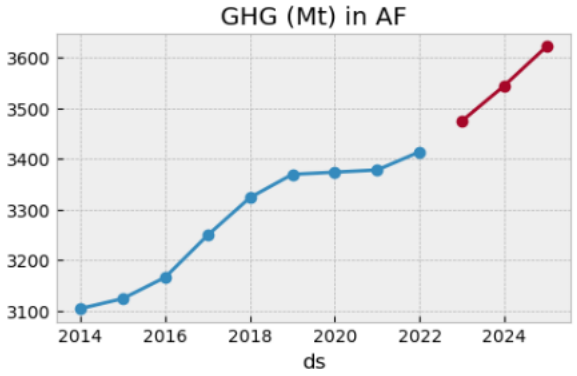
y_Agriculture	24.16
y_Buildings	-98.17
y_Fuel Exploitation	-35.33
y_Industrial Combustion	-1.19
y_Power Industry	59.31
y_Processes	9.75
y_Transport	24.64
y_Waste	16.82
population	107.70
gdp_gusd	9.26
gdp_cap	-56.34
val_pop	8.10
val_gdp_kusd	35.52
val_gdp_cap	137.30
lag1	0.69
lag2	-0.21
lag3	0.29
RM_lag1_WS3	-0.10

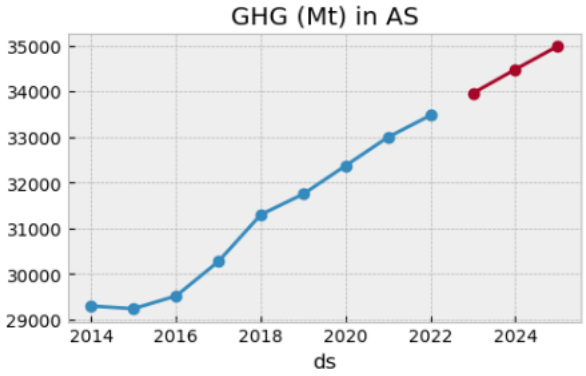
* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/3_training2.ipynb

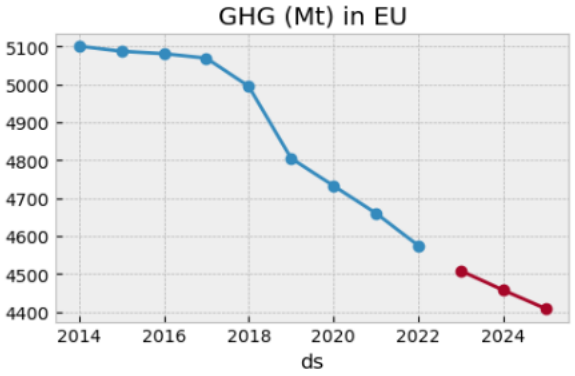
ocean

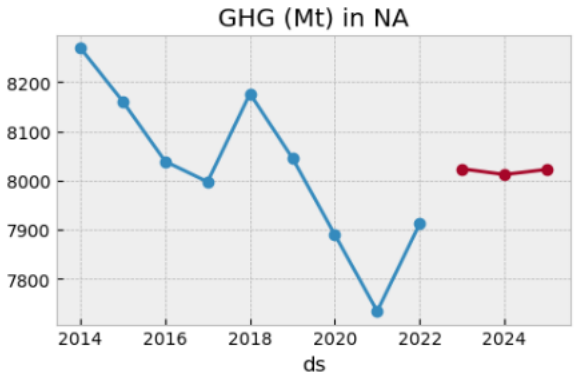
5.6. Model predictions of GHG emissions

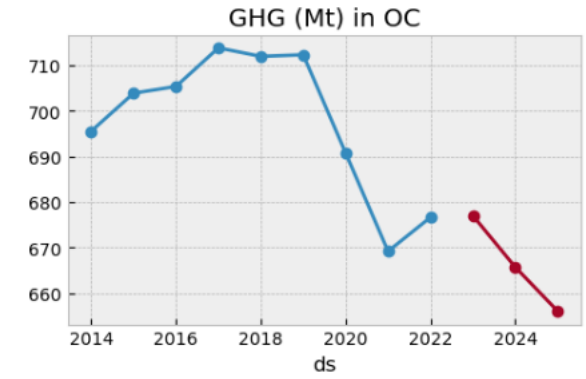
The graphs below show the 3 years of forecasts made by the best model – linear regression Elastic Net.

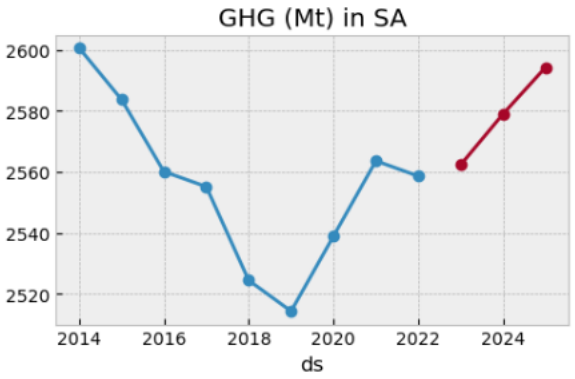












* More details you can find in the script: https://github.com/abessalov/Ocean_Climate/blob/master/3_training2.ipynb

ocean