**We don’t live CO-twice!**

**1. Introduction**

We used our data science knowledge from the Techlabs program to illustrate the extent to which co2 emissions developed in the years 1990-2014.

\*how are CO2 emissions distributed around the world?

\*Is energy consumption still that big of a deal when it comes to overall Co2 emissions?

\*ML einfügen

* **Bezug zu techlabs/ Data Science**
* **Hauptproblem - Fakten und Empfehlungen waren 2014 bereits ausgeprochen und klar → Daten zeigen absoluten Trend nach oben (Patrick)**

**2. What we did**

Like in a typical data science project, our data underwent the four main steps of data processing:

* gathering
* cleaning and sorting
* analyzing
* predicting

First, to find representative data for our analysis, we took several resources into account. As data websites like „Gapminder“ and „Our world in data“ collect all sorts of data from official institutes, it is particularly easy to find complete data sets on a variety of topics. From „Gapminder“, we obtained the statistics about energy consumption per capita, CO2 emissions per capita and the Human Development Index for all available countries since 1960.

For further analysis, these sets were put together and smoothed out. Meaning we dropped missing values and adjusted the units. During the analysis, we took a close look at the data and its behavior when combining different features. Which traits correlate? Which don’t? How did the features develop over time?

We divided the project into two parts. The first one was visualizing the data by plotting the development of CO2 emissions during the period 1990-2014 and further concentrating on interesting features as the Human Development Index and CO2 efficiency. The second part covered different machine learning approaches and their application on our dataset.

**3 Analysis with Data Science**

**3.1 Asia reaches new carbon dimensions**

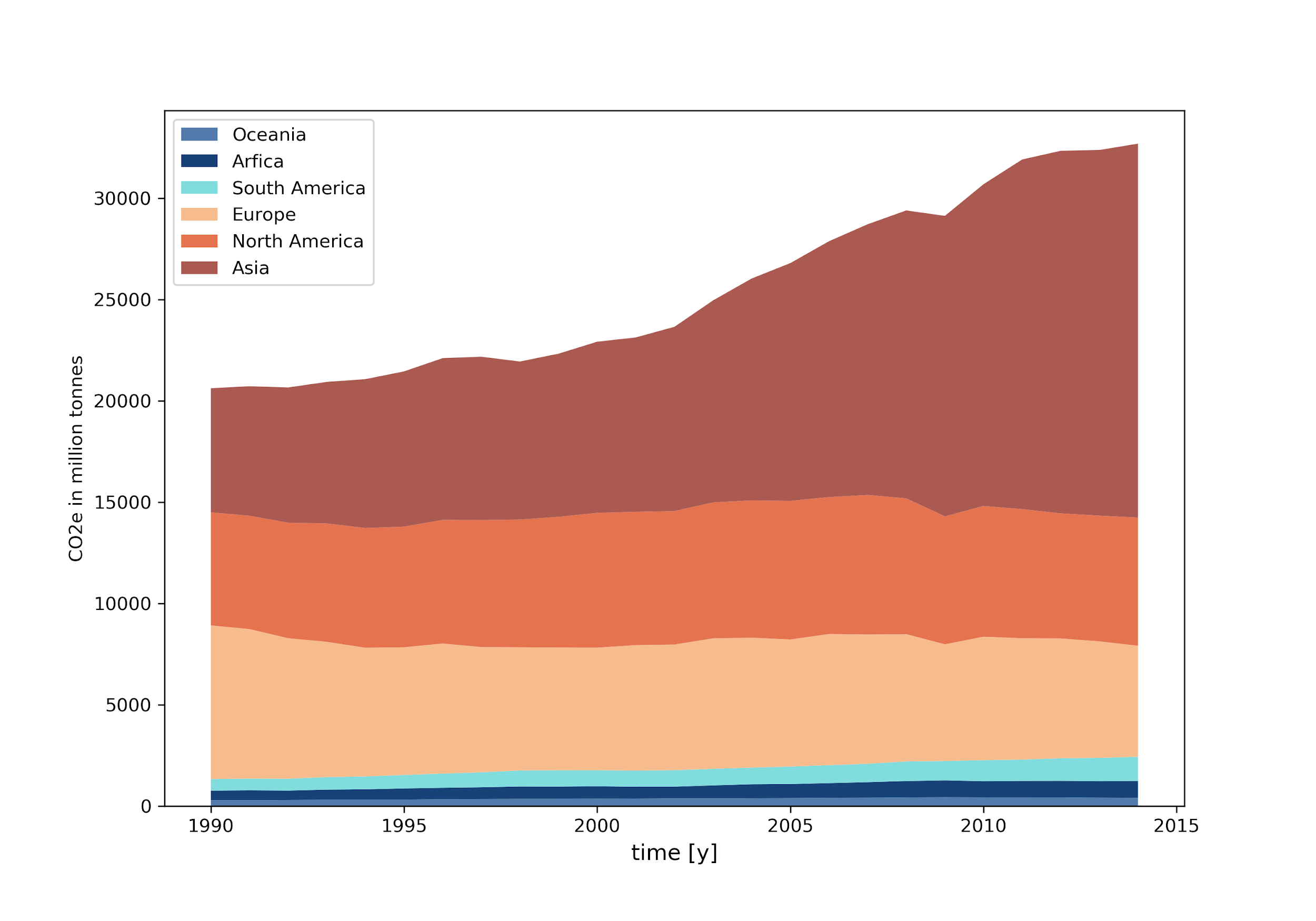
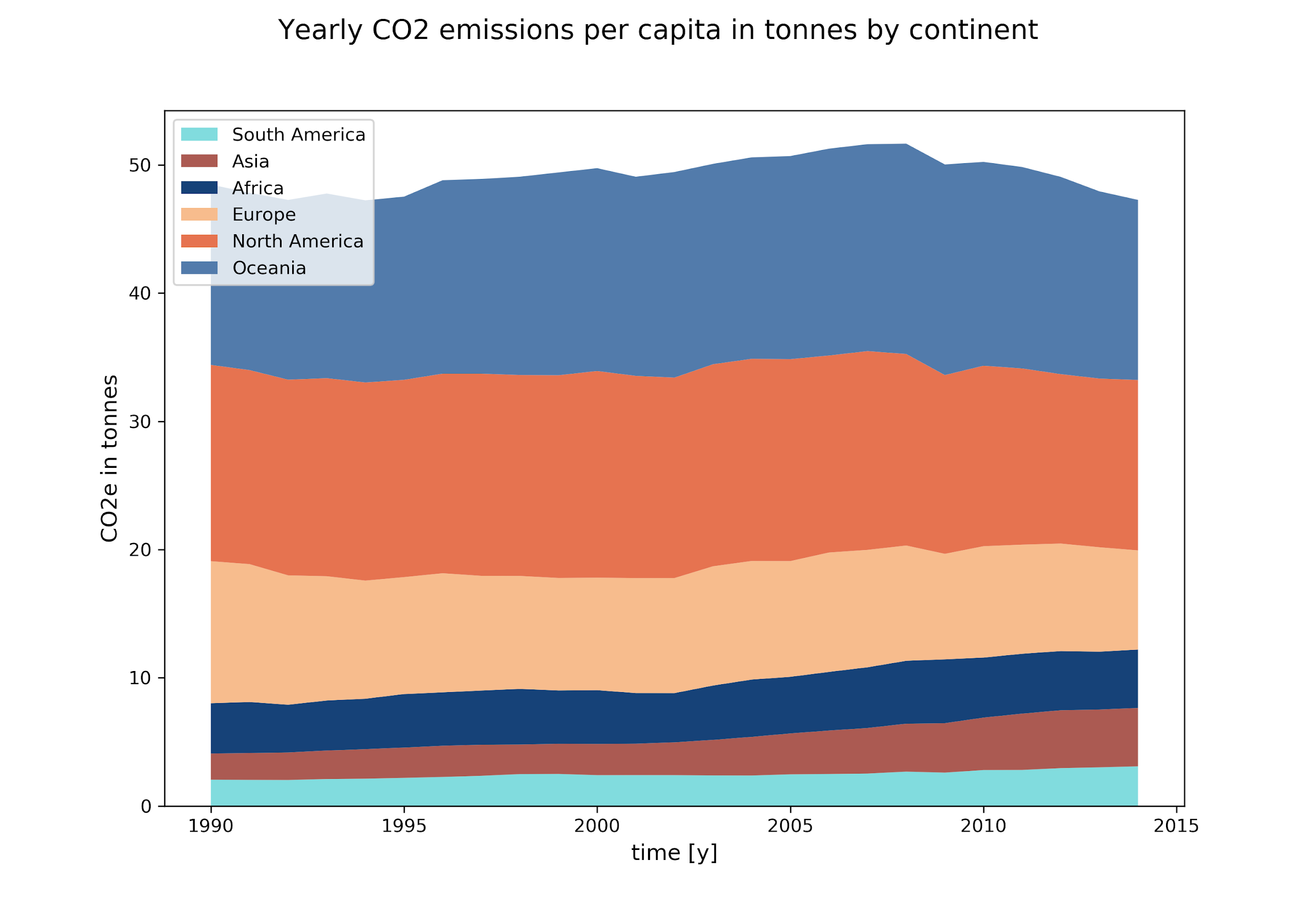
To get an idea of the CO2 emissions around the world, let’s take a short look at the emissions by continent, shown in the following plot. Up to the year 2000, the overall share of CO2 emissions was divided evenly between Asia, North America, and Europe emitting more than 80 % of the yearly overall CO2 emissions. With continuously increasing share Asia became the number one polluter by 2014, on a rising trend, with over 50 % of overall CO2 emissions. 

Figure 1: Yearly overall CO2 emissions by continent

The picture does change when looking at the CO2 emissions per capita, shown in the plot below. Asia became the continent with the highest emissions until 2014 which is mostly due to its population size since the per capita values remain secondary smallest. While this is interesting information, the main impact though can be achieved by considering the overall emissions.

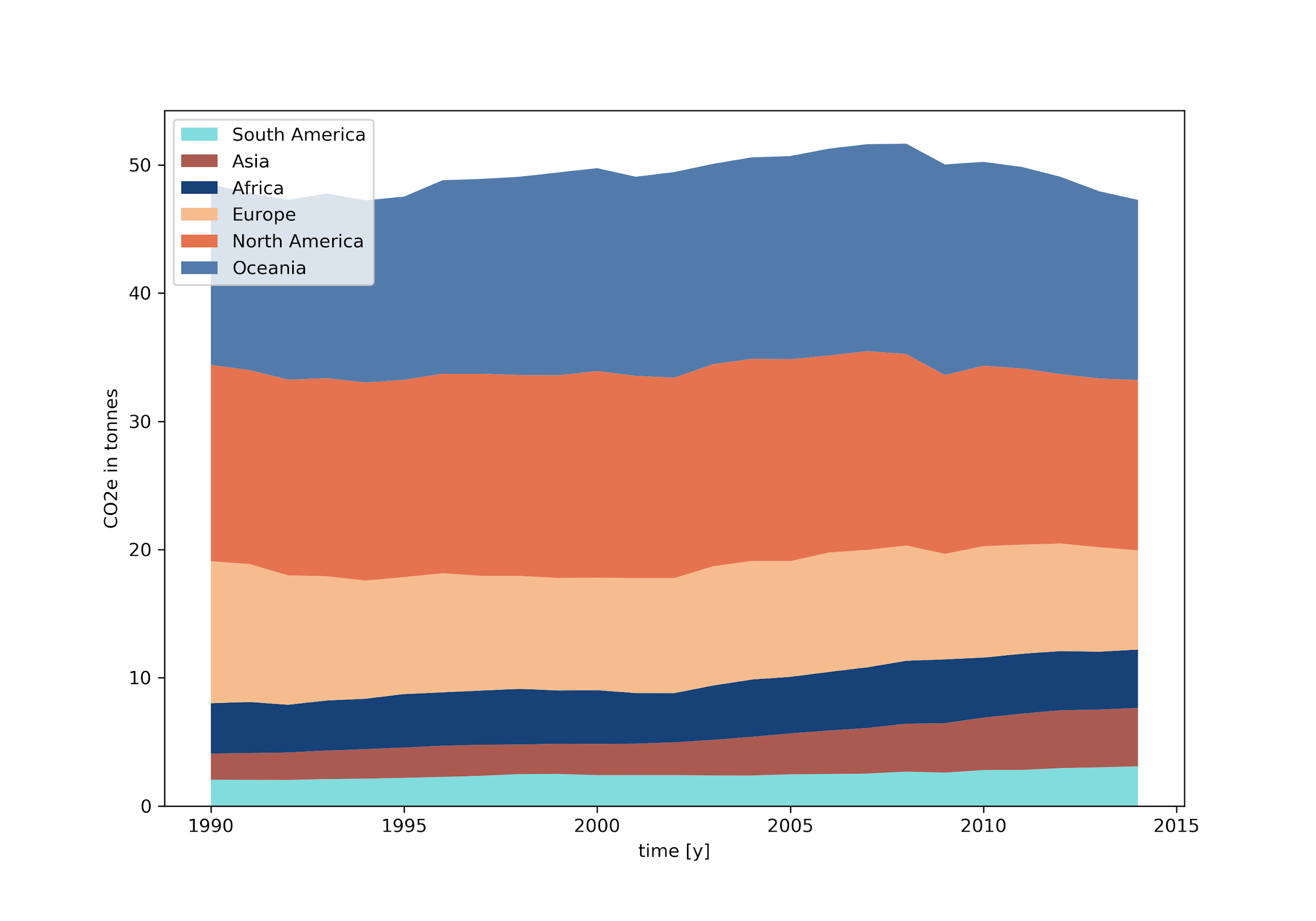
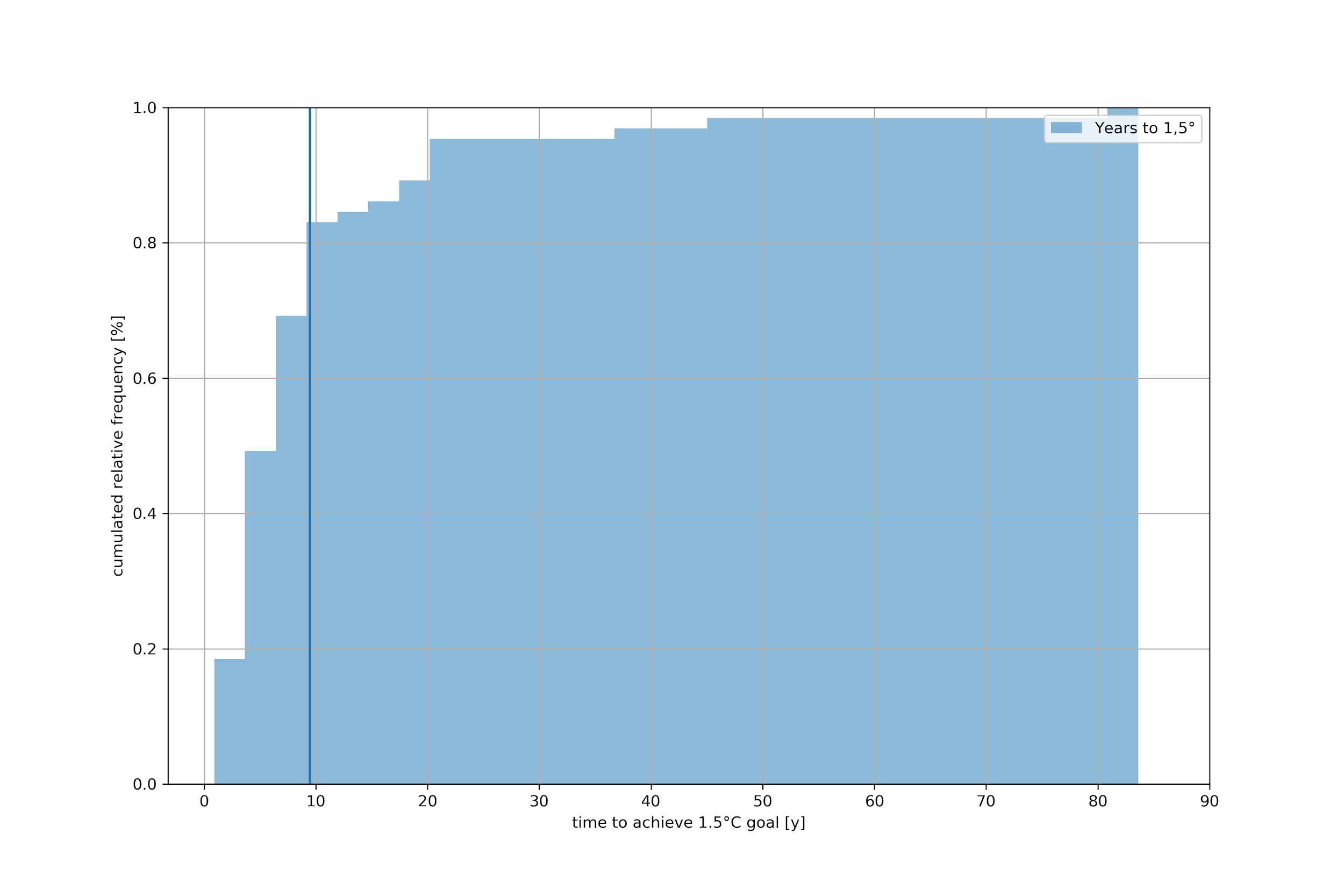
****

Figure 2: Yearly CO2 per capita emissions by continent

**3.2 How much time is left to the 1.5 °C goal?**

To meet the threat of Climate Change 189 parties ratified the Paris Agreement with one of it’s aims to limit the global temperature increase to 1.5 degrees Celsius towards the pre-industrial level. Considering the rapid rise of overall CO2 emissions, it might be almost impossible to stay under the 1.5 degree mark. By taking into account the CO2 budget calculated by MCC[[1]](#footnote-0), we determined how many years will be left until the global temperature has risen to 1.5 degrees Celsius. The result: Within a shockingly short period of 10 years, the 1.5 degree Celsius mark will be achieved if each country (considered in this study) stays at its status quo regarding the CO2 emissions at the time of 2014.

One might think the answer to this situation is easy and probably requires the big player of CO2 emissions to reduce their emission levels for the general public. The next plot was questioning this very point. For each country, we assumed that the world population is adapting the country's CO2 emissions per capita – thus: everyone emits as much as the observed nation does. In the plot below, you can see the summarized result. The number of countries (expressed as a percentage), shown on the Y-axis and corresponding years left until the global temperature has risen by 1.5 degrees Celsius, on the X-axis. The standard of living of more than 80 % of observed countries would lead the global temperature to increase by 1.5 degrees Celsius in less than 25 years with all its consequences. Even if the world population reduced the CO2 emissions per capita to the level of the least emitting county, the consequences of global warming will hit the planet in no longer than 85 years.

Figure 3: Histogram - years left until 1.5°C goal

**3.3 The impacts of climate change at 1.5 °C**

The Intergovernmental Panel on Climate Change (IPCC) Special Report on Global Warming highlights climate impacts at the current ~1 °C global warming as well as the risks of reaching a 1.5 °C and the irreversible losses that would take place at 2 °C or more warming.

To be aware of the dramatic consequences of global warming, here are just a few examples:

* Melting of the Greenland ice sheet and rising sea levels

-with the complete melting of the Greenland ice sheet, researchers expect sea levels to rise by up to 7 meters, which would cause extremely serious effects in coastal areas.

* Desiccation and collapse of the Amazon rainforest

-the global consequences would be a massive increase in atmospheric CO2 concentrations and thus a considerable increase in global warming

* Thawing the permafrost, releasing methane and carbon dioxide

-the methane and CO2 emissions from thawing permafrost soils add to the anthropogenic greenhouse gas emissions and intensify global warming. This process represents important positive feedback (reinforcing effect) in the climate system

**3.4 Emissions per capita keep climbing**

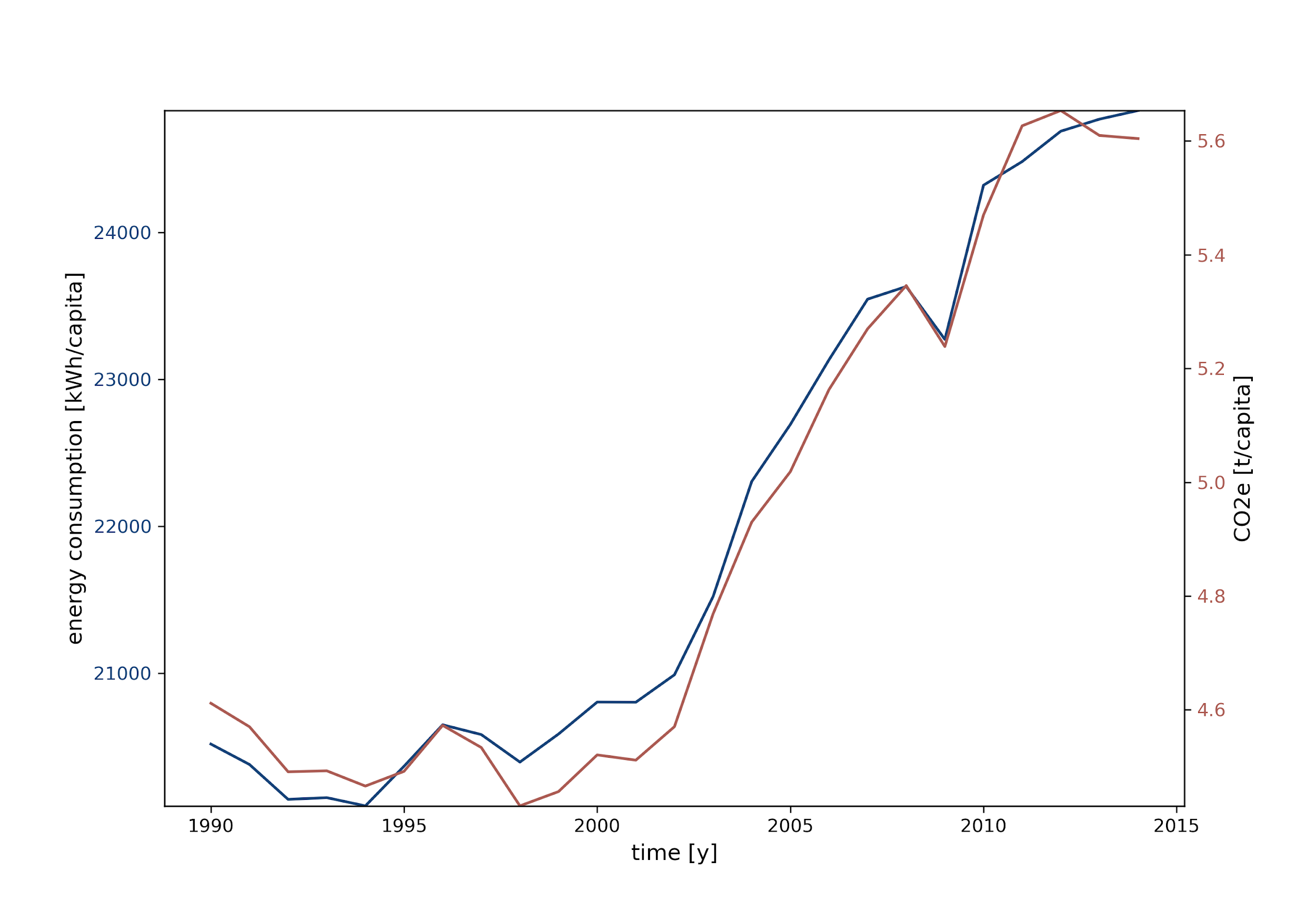


Figure 4: Comparison of energy consumption and CO2 emissions

In general, the plot shows that an increase in energy consumption goes along with an increase in CO2 emissions. At the beginning of the early 2000s, energy consumption grew dramatically and so did CO2 emissions almost equally. So do we have to limit our energy consumption? Is an improvement in CO2 efficiency enough?

**3.5 Increase in CO2 efficiency**

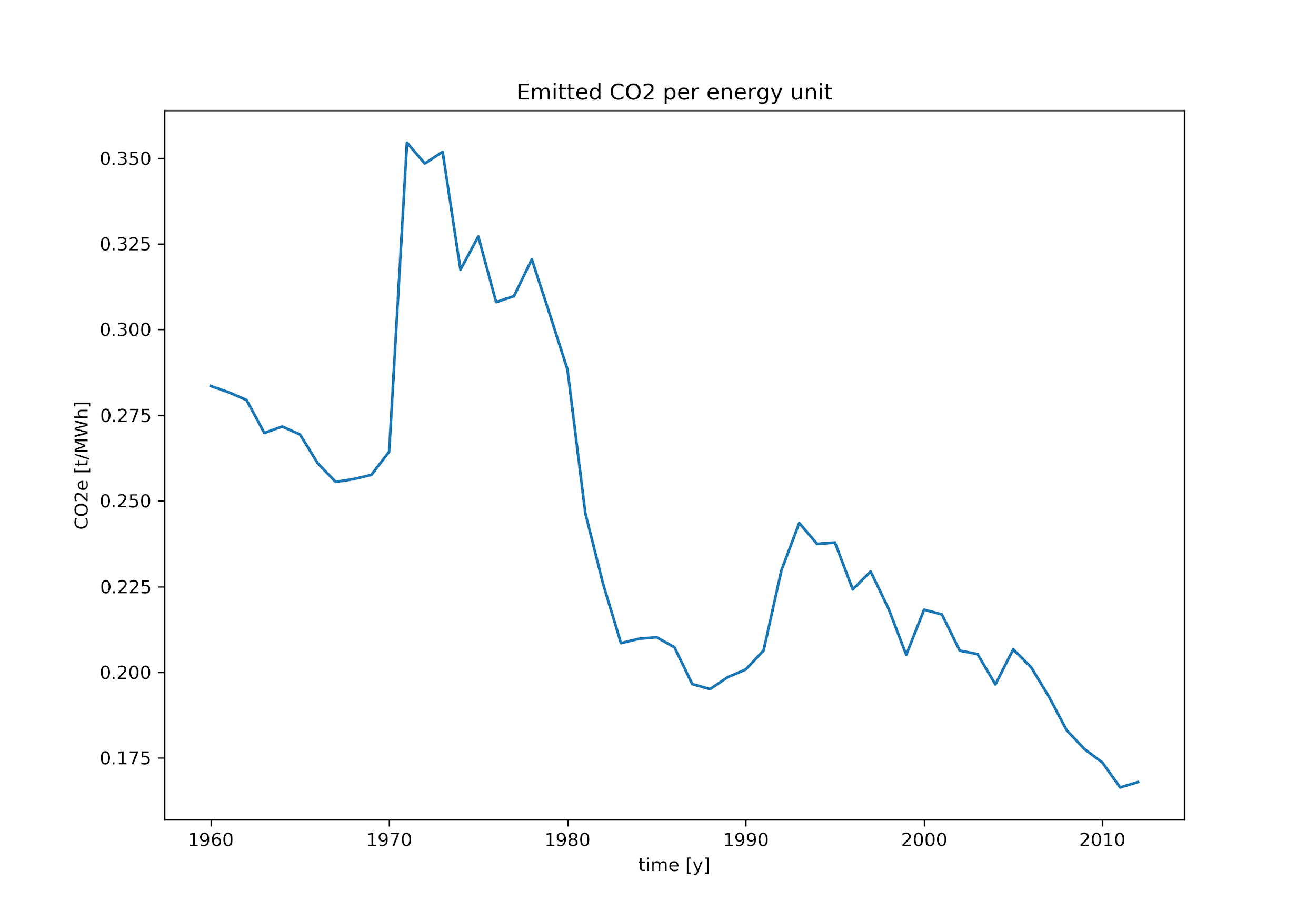
****

figure 5: Emitted CO2 per energy unit

In a perfect scenario, 1 MWh of consumed energy leads to 0 g of emitted CO2. Apparently, this has not been the case at any time, but surely is a goal to aim for as time goes on. So how did the world’s CO2 efficiency develop over the last 50 decades?

To calculate the global CO2 efficiency since 1960, the energy consumption was plotted against the CO2 emissions and a linear regression model was fitted to the values for each year. The resulting slopes were then drawn to their corresponding year.

The CO2 emissions per megawatt-hour of all nations shrank about 36% since 1960. During the 60s, the average mass of CO2 released for each megawatt-hour energy was 275 kg (606 lbs), while approximately 175 kg (386 lbs) were emitted in 2010. Iceland portrayed itself as a country with extremely high CO2 efficiency: It produced 24 kg CO2/MWh in 2010.

Fewer CO2 per energy unit is a decent step in the right direction. But it will do nothing for us if the overall emissions keep growing constantly. But do we even have a chance in reaching that goal as countries trying to achieve a higher prosperity level? Do we have to give up part of our wealth to decrease CO2 emissions in the long-term?

**3.6 Does rising prosperity always correlate with higher CO2 emissions?**

Rising prosperity is usually based on increasing economic performance and is therefore often accompanied by rising resource consumption and high CO2 emissions. In fact, the average Human Development Index (HDI) rose by 17.8% in the years under review, which is similar to the rise of the per capita CO2 emissions in the same period of time (21.6%). But does rising prosperity always come along with rising CO2 emissions?

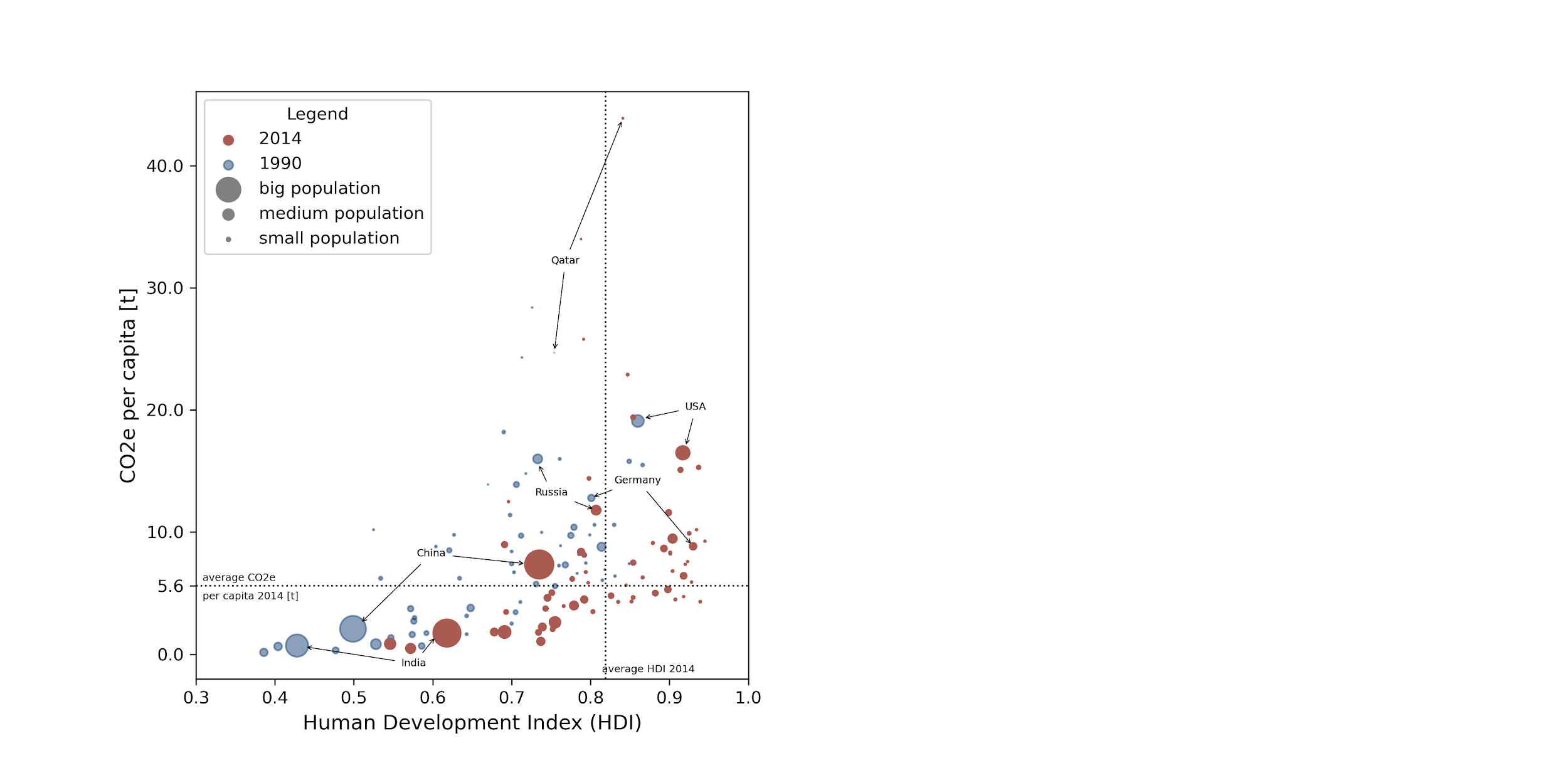


Figure 6: Correlation between CO2 emissions and HDI

The wealth of a country can be described most comprehensively by the HDI, which is made up of purchasing power, life expectancy, and education. The graph shows a rising linear trend for both years 2014 and 1990, which generally indicates a correlation between rising HDI and rising CO2 emissions.

Nevertheless, a differentiation must be made between individual countries. Industrialized countries in particular (as shown here in the example of Germany and the USA) show a negative correlation between CO2 emissions and HDI. Emerging and developing countries such as India and China generally follow a positive correlation. The influence of different countries can also be seen clearly; Qatar has the highest per capita emission of CO2, but in a global context it is not comparable to countries like China, which have a much greater influence on global CO2 emissions due to their large population. The aim of countries should be to have a high HDI along with low CO2 emissions. It sure is possible but unfortunately, there are only a few countries with an above-average HDI and below-average CO2 emissions per capita, especially when taking into account that the world’s population is rising rapidly, too.

**3.7 First Conclusion**

In the first part of our study, we aimed to display the development of CO2 emissions around the world and work out therewith associated features. Our main points involved:

* Looking at the CO2 emissions per capita – not just a few states live beyond their means but most parts of the observed countries do.
* Though the CO2 efficiency was improved it could not level out the overall CO2 emissions
* The relationship between HDI and CO2 emissions shows that at the time of 2014 only a few nations accomplished a desirable combination of low CO2-level and high HDI.

Ignoring the urgent advice of the last decades already led us to impacts as exemplarily in 2019:

Floods – abnormally high precipitation in the regions of Central USA, Northern Canada, Northern Russia, and Southwest Asia and a major South America with losses in Argentina and Uruguay estimated at US$2.5 billion.

Heatwaves - Two major heatwaves in Europe in late June and late July 2019. National records were set in Germany (42.6°C), the Netherlands (40.7°C), Belgium (41.8°C), Luxembourg (40.8°C), and the United Kingdom (38.7°C), with the heat also extending into the Nordic countries, where Helsinki had its highest temperature on record (33.2°C on 28 July).

Climate-related risks and impacts – Deteriorating food security due to the delay of the start of the seasonal rains and extremely dry periods in Southern Africa. Associated with the worst flooding in a decade affecting some parts of Afghanistan in March 2019, 13.5 million people are food insecure in the country, with 22 out of 34 provinces still recovering from severe drought conditions faced in 2018.

During the upcoming climate crisis, the interpretation and prediction of environmental data will be crucial for humanity to cope with extreme weather conditions. Thanks to uprising computational power, machine learning became a serious possibility to predict overall trends in such data.

LINK

**4 Machine Learning and environmental data (Processing environmental data with Machine Learning)**

**4.1 Overview**

Roughly, machine learning can be divided into four different techniques: supervised machine learning, unsupervised machine learning, reinforcement learning, and deep learning. For our project, we focused on supervised machine learning and compared different learning models. Using „energy use“ as the only feature variable, the models were supposed to predict our target variable „CO2 emissions“ as precisely as possible. To achieve the highest precision of the learning approaches, we fine-tuned them in the end.

The interpretation and possible prediction of new environmental data in the future is especially for a stronger climate changing world essentiallyssential. Recent trends including new upcoming, global crises make further prediction for CO2 output hard to predict for scientists, journalists, and political leaders. Therefore human computation power is limited due to physical and biological aspects. Are we able to get assistance to deal with large data or even with limited data from a different than human-based source?

With the uprising of larger memory, stronger computational power, data accessibility, and cloud computing machine learning approaches are getting more and more popular (particularly since major breakthroughs were made since the last decade`s ILSVRC, a scientific image classification competition). There are different kinds of techniques to choose from: supervised machine learning, unsupervised machine learning, reinforcement learning, and deep learning (e.g convolutional networks).

We will focus on different supervised machine learning models and their performance on two features “Energy use per capita in kWh” and “Carbon dioxide consumption per capita (CO2e per capita)”. By using only one base feature (Energy use) we will show that even with limited base features it is possible to get a good prediction for CO2 output. Further explanation will follow in the section below. Therefore some best-practices to tune the models will be presented. In the end we will summarize our results, emphasize the relevance for future predictions on the chosen features, and look at some advantages and disadvantages of supervised machine learning compared to deep learning models.

**4.2 What the relation of the used data is**

As shown above there is a middle to strong positive correlation for most of the features including mainly different kinds of power consumption, development, and CO2 consumption. Take again a short look on the correlation map, middle correlation is highlighted in lighter colors and strong correlation in darker colors

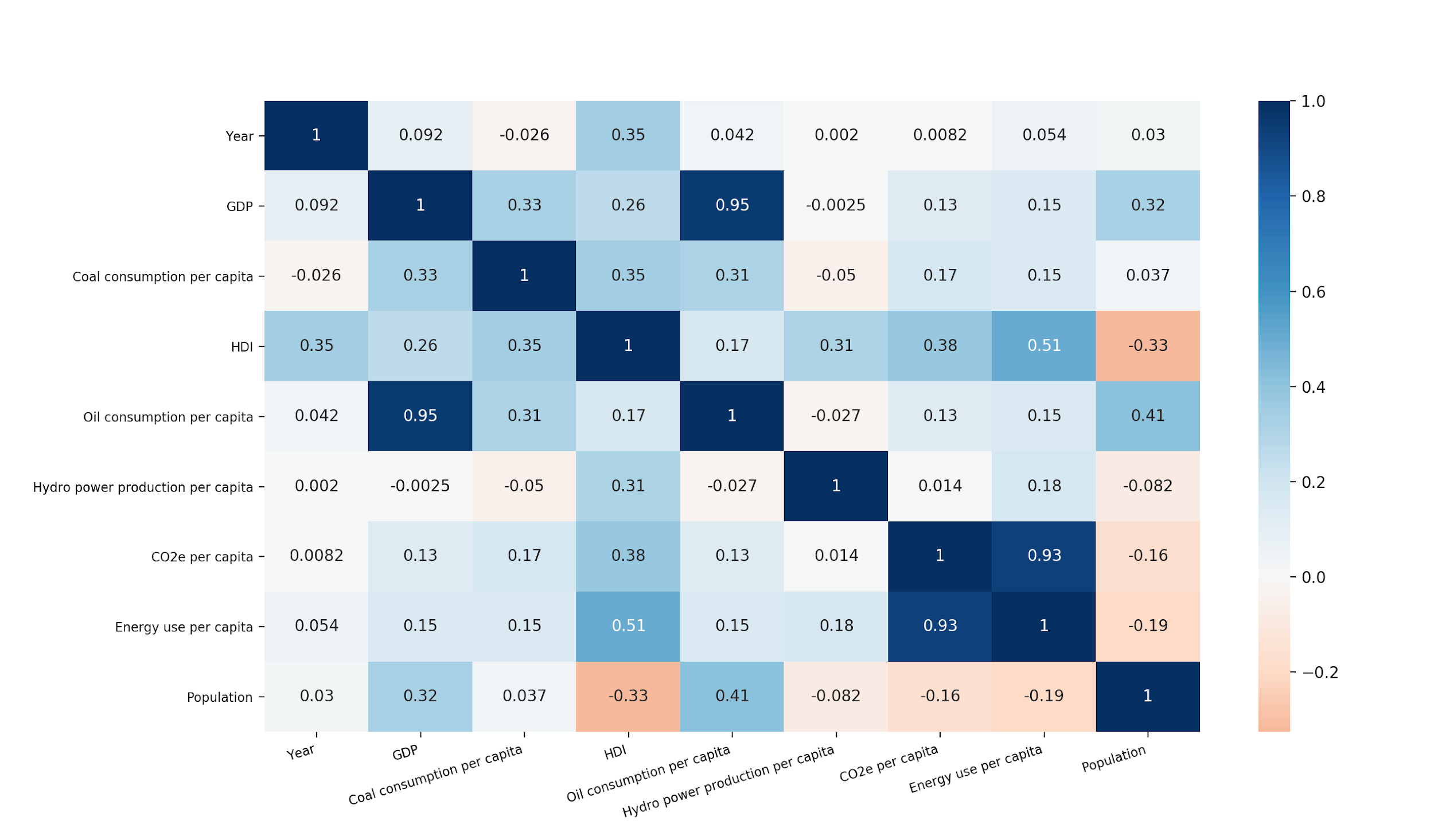


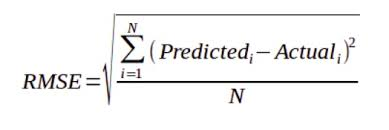
Figure 7: spearman-correlation of all used features

**4.3 Supervised models in charge**

Machine learning models distinguish between classification and regression problems. Regression problems mainly focus on continuous data while classification problems focus on discrete data. Most of the models do have a Classifier and a Regressor. As we use continuous data, we will handle our problems by using different Regressors. Furthermore we will use the scikit - learn library, Pandas, Numpy and its functions. Seaborn and Matplotlib will be used as libraries for making plots. Though there are also different libraries like Tensorflow, PyTorch, Keras and more worth checking out if one works within the maschine learning field.

It would be evident trying to use linear machine learning models to train on and predict, seeing the strong linear correlation between those features.

Now consider data matched by a model as train data and the unmatched as new data or test data. Considering this case and the fact that our model meets fully all our existing training data, is called overfitting. In the opposite case if the line would meet our training data poorly is called underfitting. In most cases there is a trade-off in model overfitting and underfitting called bias-variance-tradeoff. The measurement of “how far the new data points from our actual data points are is called root squared error or RMSE. The following formula describes it clearly:[[2]](#footnote-1)



So we want to get a predictor model based on low RMSE and good prediction score.

Let's check how different linear and non-linear models perform in our case. We create an artificial test set without using new data, instead, we use train\_test\_split from the sklearn library. After splitting into 30% test data that we are going to use for our prediction, we eventually use StandardScaler and our preferred model type within a pipeline. This scales our values from 0 to 1 so that there won't be unit problems. Although this can be useful in some cases where units differ from the usage of different data sources, in our case, it is not necessary to use thanks to preprocessing in excel!

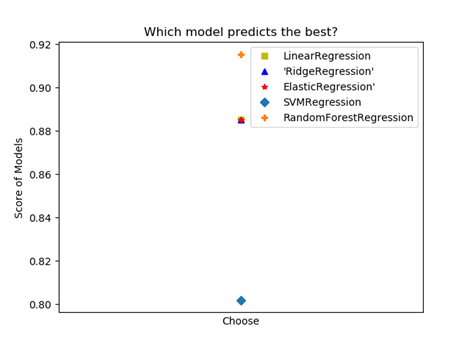
All models are tuned, so their parameters were improved by GridSearchCV (learn more in the sklearn docs). Firstly their RMSE and secondly their score on predicting new data(test data) will be shown above. Take a short look

Figure 8: Which model predicts the best?

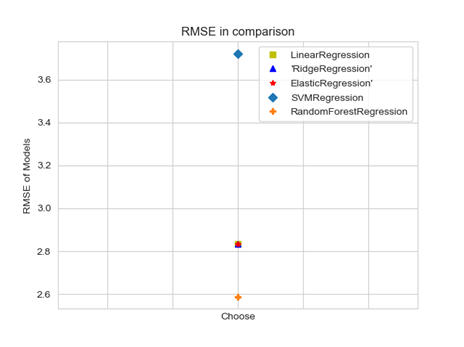
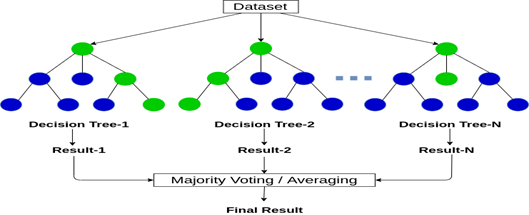


Figure 9: RMSE in comparison

Our linear regressors RidgeRegressor, LinearRegression, ElasticRegression all perform nearly evenly. In Contrast to LinearRegression, RidgeRegression, and ElasticRegression both use penalties as parameters to reduces to the slope of the model(or in this case curve) to fit better to the test data.

Although this improves the fit slightly our model it is still too simple for our testing data. In contrast, RandomForestRegression which uses bootstrap to create subsamples of our actual data and then averages through those data points to make predictions seems to perform quite better on data that has positive linear indication with some few, strong outliers.

**4.4 How RandomForestRegression works and tuning**



Abhishek Sharma, May 2020

So, let us take a look at how RandomForestRegression actually works and in which cases it would be useful to try out RandomForestRegression. The parent idea of RandomForestRegression comes from DecisionTrees, but it is expanded by the means of accuracy. After creating a subset called the bootstrap set, it takes a feature from the samples as the root node and the other features as nodes. This is done multiple times

In Regression, decisions are made by averaging the samples within a node. The overall action of bootstrapping and decision making is called bagging. Not all samples are used in the model, roundabout 1/3 is left out. This left out samples can be used instead of cross-validation. Nodes and the decision making is conditional wise based, also it tends to overfit, because of the bagging factor. Remember, we wanted a more complex and more fitting model to use, so this one is doing a perfect job.

In terms of finetuning there many parameters to choose from and that are advancing the complexity of the model.

max\_depth: Customizes to what extend (depth) one wants a tree to grow. Especially essential for datasets with many, quasi dependent features.

min\_sample\_split: established that the minimum amount of sample values within a node before the next split. Useful to get away impurity, also for customizing fitting of the model. More samples will underfit and less will likely to overfit the model.

min\_samples\_leaf: established the minimum amount of sample values within a leaf node after the split. Useful to get the fit right.n\_estimators: the use of the number of trees used. Increases the complexity of the model and the computation time of your device. CPUs using a slow amount of hertz (instruction per second) and single processing (cores of CPU) will take a significantly longer time to compute.

There are several other Parameter to use for finetuning, but those above were the most important in our case.

**4.5 Conclusion**

As seen above, machine learning can be a powerful tool to predict new data even with a few data as a basis. But not only with a few, but also with a large set of data and multiple categories of features machine learning´s performance is amazing. There are several methods out there to use and of course, deciding which to use can be difficult. But especially when it comes to that one wants to have

an overlook over the process of input and output, supervising is probably the method you are going to choose instead of unsupervised learning because there is a link between input and output (by determining your input and output data you want to work on). Algorithms used by supervised machine learning are support vector machine, neural network, linear and logistic regression, random forests, and classification trees. I recommend taking a first look at the data, think about the course of the data, and even if there is a context between. In most cases after doing this, one is able to limit the number of useful models. Further, it is important to distinguish if one´s data is continuous or discrete. For first one will use regression models, for the second one will use classification. Parameter and tuning with GridSearch or even RandomGridSearch will improve scoring, *RMSE,* and the fit of the model. Testing on the model is done by splitting the data into train data and test data by *train\_test\_split*. Lastly, modeling and predicting will help with analysis, though analysis is worthless without the interpretation of humans. So, ask yourself every time when performing certain steps: What am I doing now? Why I am doing this step? How will this affect my data? What is the context of doing machine learning on this data?

Even if we were looking mostly on the best practices of machine learning with the data given, let me express for a minute the importance of reducing not only CO2 emission of fossil energy sources but also the overall reduction of energy consumption. As seen, there is a strong correlation between CO2 emission and energy consumption. Even new data of energy consumption and CO2 emission show the same pattern as they fit 91,5% in our machine learning model. Certainly, it will be important not only to switch energy resources but also to reduce overall energy consumption (IEA (2019), Global Energy & CO2 Status Report 2019).

1. Mercator Research Institute on Global Commons and Climate Change <https://www.mcc-berlin.net/forschung/co2-budget.html> [↑](#footnote-ref-0)
2. https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2018/05/rmse.png [↑](#footnote-ref-1)