Analysis Domain, questions and plan

Identify the application domain and datasets

I aim to investigate the hypothesis that the countries affected most by the climate crisis, in terms of surface temperature, are those that have contributed the least to it, in terms of carbon emissions. I will also classify countries based on environment related characteristics such as carbon emissions and biocapacity using clustering.

Both datasets are sourced from Kaggle: the National Footprint Accounts (NFA) data were gathered by the Global Footprint Network (GFN) and the Earth Surface Temperature (EST) data by Berkeley Earth.
The NFA data contains country-level data on ecological footprint and biocapacity, over the course of several years up to 2014. The EST data contains average temperature per country for each year since 1750. There is both a country-level dataset which has average temperature for each year since 1743 (though not for every country), and a dataset aggregated across all countries.

Identify questions and analysis tasks

I found several papers containing similar data analysis conducted in the domain of climate and environmental data as to what I intend to accomplish. Mahlstein, Kautti (2009) used k-means to detect and map regional climate change patterns, moving away from a previously used rectangular shaped representation to a more representative fluid model. The clustering was based on local climate patterns and model uncertainty.
Kremer, Gunnemann, Seidl (2010) used cluster tracing (mapping similar clusters over time) in conjunction with time series analysis to detect long term vs. short term changes in climate. The analysis split time into equal length intervals, and standard clustering was applied for each period of time. They then used a sliding window approach, specifying overlapping chunks of time and analysing them in order to detect change.
Sarker, Alam, Gow (2012) used data from 1979 to 2009 at aggregate level to assess the relationship between climate variables and rice yield using ordinary least squares (for normally distributed rice yields) and median (quantile) regression (for rice yields that were not normally distributed).

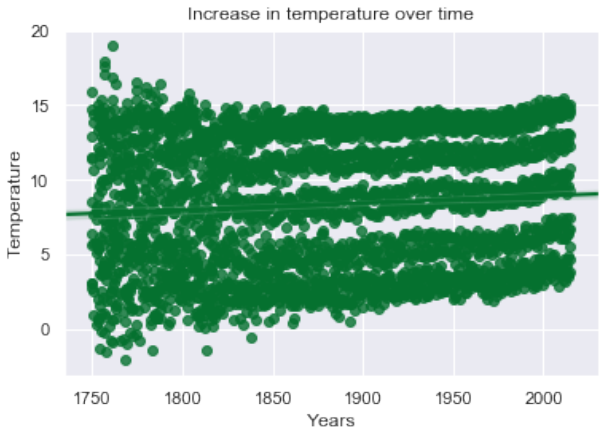
Initially explore and develop an analysis strategy

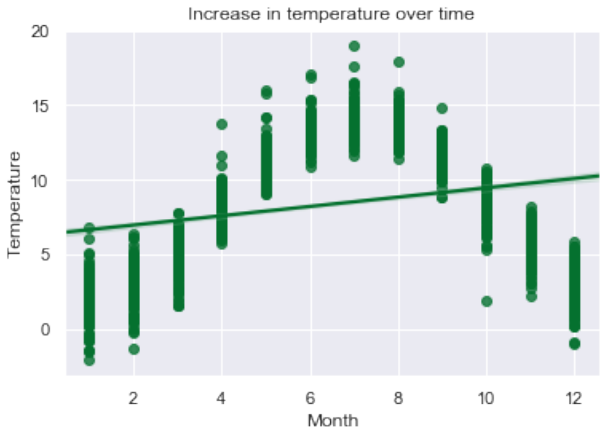
I plan to simplify the datasets separately before bringing them together based on country and year, as to join them together in their original form would result in a very large and wide dataset. I need to see how far back the data goes in terms of years for the countries and impute or remove any nulls in the data, and check for outliers. After merging, I will do some correlation analysis on the variables. I will engineer a couple of new variables – biocapacity deficit per capita and total biocapacity deficit (country-level).
I will run PCA on my data to reduce its dimensionality and cluster countries based on their carbon output/rise in temperature initially using k-means but may change the clustering algorithm if the

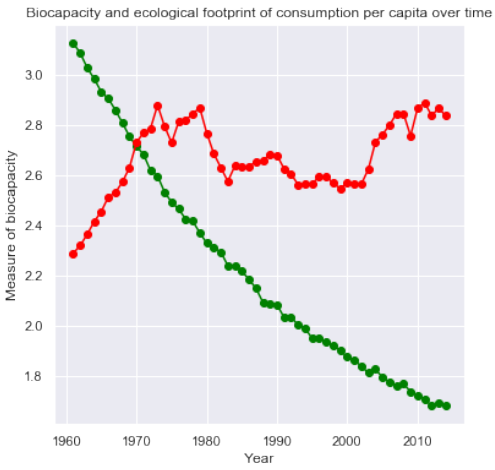
data requires. I will analyse the clusters to see how they differ from each other and what that can tell us about the countries and their relationships to each other and their average temperature and ecological footprint.

Findings and reflections

General world trend:







Between 1750 to present, the average land temperature and ecological footprint of biocapacity per capita has increased, while biocapacity per capita has decreased.

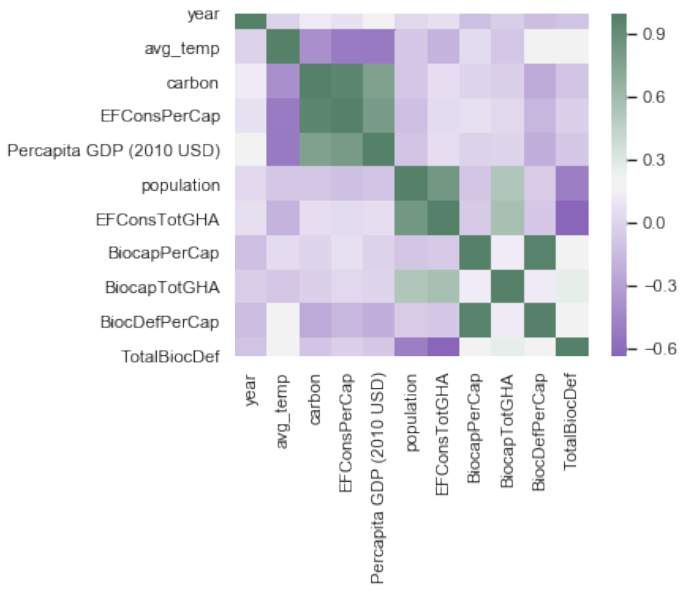
Correlations between variables

The correlation analysis I ran found strong positive correlation (>0.60) between:

- global hectares of world-average forest required to sequester carbon emissions and ecological footprint of consumption per head (0.95).
- population and ecological footprint of consumption measured in global hectares (0.84).
- global hectares of world-average forest required to sequester carbon emissions and per capita GDP (0.77).
- ecological footprint of consumption per head and per capita GDP (0.81).

Is also found strong negative correlation (< -0.60) between:

- total biocapacity deficit and ecological footprint of consumption measured in global hectares (-0.64).

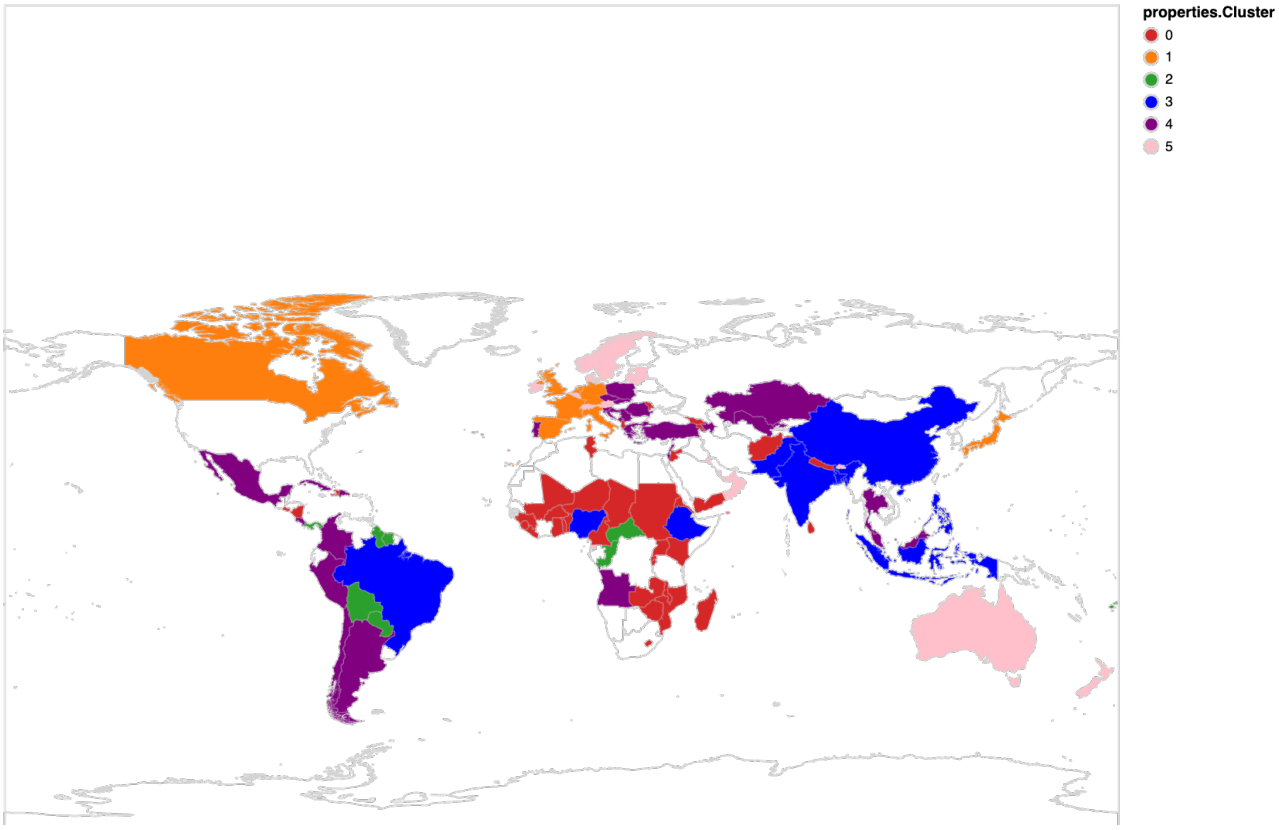


Clustering

The GMM clustering modelled 6 clusters within the data, although the cluster distribution wasn't great: the first cluster - cluster 0- contained more data points than the others, while clusters 1 & 3 were quite small. From the point of view of the categorical variables, clusters 0, 4 & 5 were representative of most UN subregions, while clusters 1, 2 & 3 were less so. The numerical characteristics of the clusters were observed in a radial chart, below.

Choropleth maps illustrated the distribution of the clusters across the world, over time. Over the time span of 1961 – 2011, they showed a relatively stagnant picture in terms of which countries fell into which clusters, however there were some excepts, e.g Canada moved from cluster 1 to cluster 5 between 1961 and 1971, more countries in Africa fell into cluster 0 over time. There are obvious overall geographical differences, China, India, Indonesia, Malaysia and Pakistan are all in cluster 3, Western Europe and Canada fall into cluster 1, Australia, New Zealand and also most Scandinavian countries fall into cluster 5.

2011



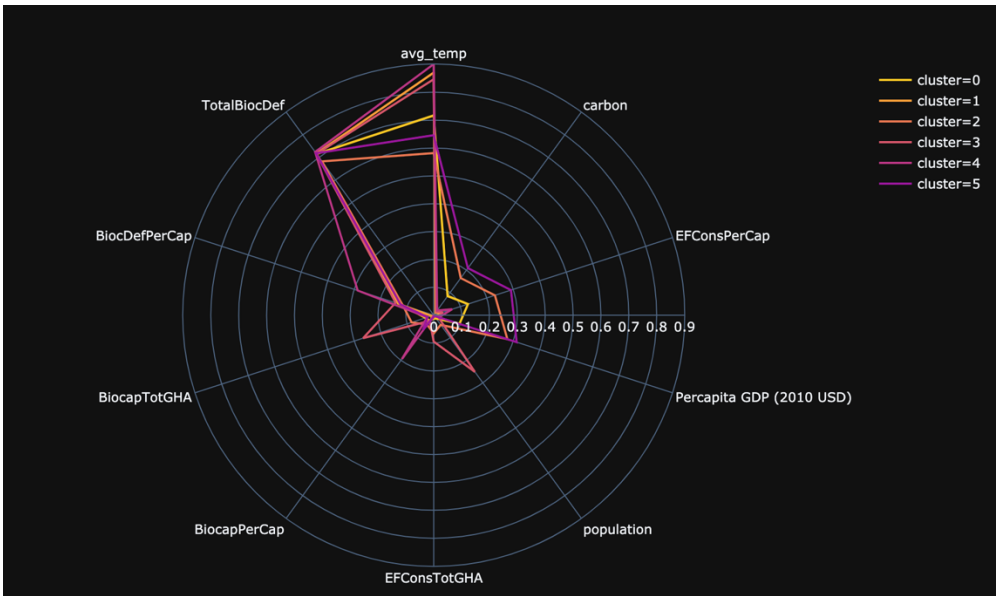
Radial chart

When I initially ran the clustering I found a lot of countries I didn't expect to fell into the 'high carbon emissions' cluster – this was due to my imputation of null values in the data cleaning step in the carbon emissions and GDP columns. I decided to rerun the analysis and just remove the

nulls instead of imputing them, leaving fewer countries to be studied, but a more accurate view of them.

The radial chart shows us the differences between the clusters in terms of their numerical characteristics. I have also listed some of the countries associated with those clusters (in 2011):

- **Cluster 0:** high average temperature and total biocapacity deficiency, low total biocapacity deficiency per capita, negligible all other variables. Central and south Africa, Afghanistan, Nepal, Haiti, El Salvador, Nicaragua.
- **Cluster 1:** medium average temperature and total biocapacity deficiency, medium carbon, GDP per capita, ecological footprint of consumption, low biocapacity in hectares, negligible all others. Northern Europe, Canada, Japan.
- **Cluster 2:** high average temperature and total biocapacity deficiency, medium bio deficiency per capita, bio capacity per capita, negligible all others. Bolivia, Paraguay, Panama, Central African Republic.
- **Cluster 3:** high average temperature and total biocapacity deficiency, low biocapacity deficiency per capita and ecological footprint of consumption, medium population, biocapacity in hectares. China, India, Pakistan, Indonesia, Philippines, Brasil.
- **Cluster 4:** high total biocapacity deficiency, less-high average temperature, low total biocapacity deficiency per capita, ecological footprint of consumption per capita, GDP. Mexico, western countries in South America, Eastern Europe, Cuba, Thailand, Angola, Malaysia.
- **Cluster 5:** high total biocapacity deficiency, less-high average temperature, medium carbon emissions, ecological footprint of consumption per capita, GDP, low biocapacity deficiency per capita. Ireland, Australia, Scandinavia, New Zealand, Oman, Switzerland, Belgium.



There were a number of limitations on the analysis. There were few countries with comprehensive data: the choropleth maps ended up showing 115 countries, due to the presence of 162 countries in the NFA data and 242 countries in the EST data, and not all of them overlapping in terms of the

inner join. This could also be because the data was joined on country name, instead of a unique country code, which may have not accounted for differences in nomenclature, spelling or punctuation. The country-level temperature data was only complete enough to look at after 1961. I would particularly have liked to have seen which clusters the U.S., Russia, Ukraine and Greenland would have fallen into, and whether their impact would have affected the clusters.

In terms of my findings, I am not sure how much valuable insight I have managed to gain from my analysis. Going into the project I was intrigued to look into the patterns between ecological footprint and average temperature but haven't managed to find anything conclusive. This was probably to be expected considering the limited scope of the data. Clusters 0 and 3 both experience high temperatures while having low carbon emissions, while countries in cluster 5 have lower temperatures and higher emissions, showing clear distinctions between countries and their impact and the effect of climate change on them. However, the radial chart doesn't show as much difference between the clusters as I would have expected and the clusters all have a similar shape on the graph. Were more conclusive findings to be found, the insight gleaned from this data could be used to classify countries based on their relation to the climate crisis, and tailor how different groups of countries should respond in order to efficiently resolve it.

References

1. Mahlstein, Knutti 'Regional climate change patterns identified by cluster analysis', 2009
2. Kremer, Gunnemann, Seidl, 'Detecting Climate Change in Multivariate Time Series Data by Novel Clustering & Cluster Tracing Techniques', 2010.
3. Sarker, Alam, Gow, 'Exploring the relationship between climate change and rice yield in Bangladesh: an analysis of time series data', 2012.

Datasets used:

- https://www.kaggle.com/footprintnetwork/national-footprint-accounts-2018
- https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data
- https://hub.arcgis.com/datasets/a21fdb46d23e4ef896f31475217cbb08_1 (shape file)