COUNTRY EMISSIONS PREDICTION- TIME SERIES 1990-2019

Our goal is to build a predictive model for each country future emissions.

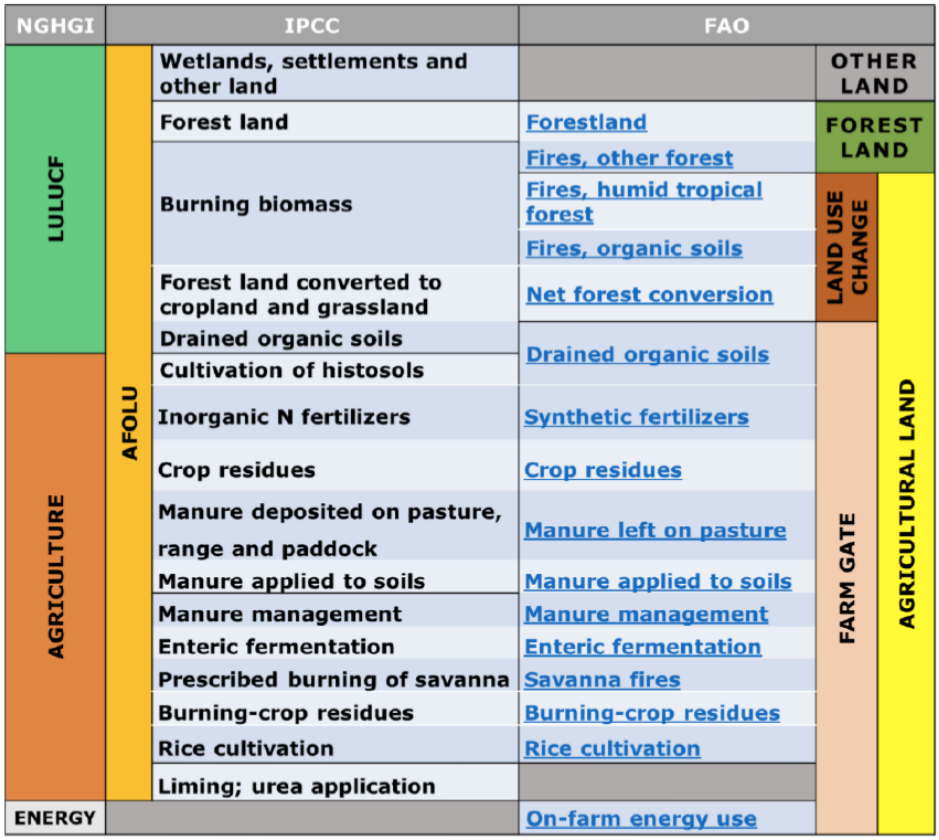
We will work with total emissions datasets downloaded from FAOSTAT, https://www.fao.org/faostat/en/#data.

1. Data Exploration

The original dataset at FAOSTAT includes the historical series from 1961 to 2019, with two emissions predictions for 2030 and 2050. We will filter this and work with the final series 1990-2019. This way we will have most of the countries with data, or at least with enough years data to build an accurate predictive model.

The emissions values that will be used in order to get each country total emissions will be "Emissions (CO2eq)". More specifically, values for "Farm-gate Emissions" and "Land Use change". These two categories inside Emissions take in consideration the biggest part of the emissions created during food production, as stated in FAO resources, https://www.fao.org/documents/card/en/c/cb5293en. As a detail of this see next:

Figure 13. Correspondence between NGHGI, IPCC, FAO Land Use and FAOSTAT emissions categories



Source: FAOSTAT, 2021.

The unit value for emissions is kilotonnes.

2. New features

In order to add more data and features to the emissions dataset, we will merge this with next FAO datasets: population, food production and GDP. In all three merges we will continue framing the dataset into the series 1990-2019.

As a result of the new dataset, we will generate the next new features:

- Population by country, value unit x1000 people. New generated features are:
    - Total Population.
    - % of world population.
    - Population anual growth.

- Production by country. We will keep info under tonnes unit value. New generated features are:
    - Total production.
    - Production per capita (country).
    - Production per capita (world).
    - Production anual growth.

- GDP by country. New generated features are:
    - GDP total.
    - GDP per capita (country).
    - GDP per capita (world).
    - GDP anual growth.
    - GDP anual growth per capita.

- Emissions. New generated features are:
    - Total Emissions.
    - % world total emissions.
    - Emissions per capita (country).
    - Emissions per capita (world).
    - Emissions per GDP unity (country).
    - Emissions per GDP unity (world).
    - Emissions per production unity.
    - Emissions anual growth.

3. EDA and graphics

To have a better understanding of the final dataset, we will generate different plots, with plotly express. Some of them are also present at the clustering model, since in essence we are working on the same final dataset. Some of these plots are:

- Scatter by country, all years included, color filtered by continents. This will provide a static picture for each country.
- Scatter by country, all years included, color filtered by continents. Detail of America continent.
- Histogram for total emissions, with country population size. Can be filtered by continent or continents.
- Line plot by continent and/or country.

4. Dataset final cleansing.

- Country Code. We will add this feature to each country, so that we can have all countries under a numeric value, instead of being a categorical feature. This will make easy the process of passing each country to the XGBoost, plus will keep it simple to recover the country name after the model has been created.

- Use of KNN Imputer for null values.

5. Create new features (time series based)

- This will get done with different Group By with the date feature(Year), so that we can generate new features to improve our forecast/predictive model.
- After different Group by trials, we will keep the ones that generate time related features that add more value to the final model (smallest possible RSME). This was done under an iteration logic, with a final understanding on what are the features that when combined help to better predict future countries emissions.

6. Before model training

The final dataset to be trained will be generated by selecting which features will be passed, and partitioning the dataset in train, validation and set. 2018 will be used for validation, and 2019 for test. This way we could have some final metrics on success prediction, since we will have a predicted emissions value for each country for 2019, and at the same time we do have the real emissions value at our original dataset before training.

7. Training with XGBoost

RMSE will be our measure of model success for prediction.

8.Feature importance

This function will help us to visualize which of the generated features have helped to better predict the emissions, whether they are features created by ourselves when merging datasets, or the features generated under time series logics.

9. Prediction and evaluation of the model

In order to evaluate the success of our final best model (lowest RMSE in test), we will see the deviation between predicted and real emissions values.

Success criteria will be established at 5% variation (predicted value versus real value). Any country with a variation bigger than 5% will be considered a failed prediction, and opposite way for any variation smaller than 5%.

10.Results

The final forecast/prediction for emissions achieved to predict 120 countries emissions with a variation from real values smaller than 5%. These 120 countries explain together a 96% of the world emissions. If the success criteria is lowered to a 2% variations of the predicted value from the original one, then we can see that 89 countries are included, explaining a 89% of the world emissions.

There is a group of 90 countries with predictions variation over 5%. When profiling those countries, we can see that are all small countries, and a big part of these are islands. There is probably some room for improvement into the generated data for these countries.