VISUALIZING GHG EMISSIONS, 2015-2021

Documentation detailing data preprocessing & technical details

Executive summary

Climate change is a critical challenge we must address as urgently as possible. As COP28 started, I wanted to create a simple but informative dashboard visualizing for everyone overall and country-level emissions to drive home the message: we're not on track. The dashboard is accessible here. This short document details how the data was acquired, processed, and used.

Data sources

The project mainly relies on the Climate TRACE dataset that, as of November 2023, contained information about country-level and sector-subsector level emissions of carbon dioxide, methane and nitrous oxide between 2015-2021. This dataset was directly downloaded from the data download section of the website. As of December 2023, this option is no longer available to download all the data (yearly overviews can still be accessed), but the entire dataset and more is still accessible through the Climate TRACE API.

The other data source for the project was the Gapminder population dataset so that yearly accurate population data could be matched to the different countries. The V7 version was accessed and downloaded from the Gapminder website. This was filtered to the same period (2015-2021). Both data sources are under a Creative Commons 4.0 License, which means anyone can use or distribute it for any purpose while acknowledging the source.

Climate TRACE is an interesting data source as it's an independent emissions monitoring provider whose data is freely accessible. Their data collection relies on satellite and remote sensing data, and they use advanced processing and artificial intelligence to provide site-level estimates globally. As using site-level data was beyond the scope and purpose of this project, I opted for country and sector level data. For the period 2015-2021, this meant 15 852 observations for 9 sectors and 252 countries or territories for the before mentioned three gas types.

The potential issue with this data is that it relies on AI algorithms to provide estimates for sites that cannot provide third-party verified reliable emissions information, combined with satellite data. They leverage 'ground truth data' for training, but any biases inherent to that verified dataset (such as a certain type of emission being less

measurable than others and thus not well represented in the verified dataset) might be also part of the final estimates.

Another potential concern can be privacy, as Climate TRACE collects information about the location and impact of all polluting activities. At the same time, if anything, this seems like a moral ambition and I would find it hard to argue against collecting secret emissions data. In case of military facilities, or for individuals living in poverty causing them to emit more (e.g., cheaper but worse fuel sources), there might be some privacy concerns worth considering. But as of now, this is one of the most detailed emissions databases that is freely accessible and contains data on 350 million individual emitting points or territories, making it a great data source to work with. Also, at the aggregated level I'm working with, these privacy concerns are not yet that relevant.

Data preprocessing

As the Climate TRACE dataset didn't contain shape files or full country names, first I used the R Naturalearth dataset to add country names and shapes to the data. Then, in line with observed practice, I divided all emissions data to be able to display it in million tons instead of tons. The Naturalearth and Gapminder datasets had different country naming conventions in 14 cases, and I modified the Gapminder data to align with the Naturalearth names (e.g., Slovak Republic to Slovakia).

I merged the datasets on year and country. Lastly, for 2021, I also calculated per capita emissions on a country level, converting the values back to metric tons. I noticed a few countries, most notably the Bahamas and Panama with outstanding per capita emissions. This is probably caused by an underlying mistake in the dataset, one that I will reach out about to the Climate TRACE team. In the meantime, I substituted the amounts with data from Our World in Data. For the rest of the countries, even if the TRACE data indicated higher emissions, I left their per capita values as they mostly align with other data sources.

Visualization and dashboard elements

To visualize and analyze the data (as well as for all preprocessing steps above described) the R language was used. Specific packages leveraged were mainly the `tidyverse` for processing and merging, as well as `sf` and `rnaturalearth` for some spatial aspects. Visualizations were created with `ggplot` and `plotly`, and the `viridis` and `RColorBrewer` packages were used for additional color options. Finally, `Shiny` was used as the dashboard platform. Shiny is an openly accessible online platform that can be used to publish web apps using only R or Python. The dashboard is published on the freely available Shiny server.

For the visualizations themselves, I had to further transform the data a couple of times. The first page shows cumulative emissions per gas type, here the data was

summarized on a country, year level. For the map view, I used total and per capita emissions without transformation for the displayed values, but for the total emissions, I used log-scaled values to produce the choropleth map as China and the United States emit so much the rest of the world was not well visualized next to them (this is noted on the dashboard as it's an important change). For the second and third page, similarly, the data was summarized on a country – year level. For the 2nd page, countries were ranked based on their total emissions in the given time period in a given gas, and the top 5, 10, and 15 emitters are highlighted. For the second chart of the 3rd page, the data is summarized on a year – country – sector level.

All of the technical documents (preprocessing, visualization, and final data) are available in this public GitHub repo.

Dashboard results and next steps

This dashboard was created as a thought-provoking reflection on COP28. While Climate TRACE also visualizes and explains this data in much more detail, I wanted to provide a simple and easily understandable picture. The findings are simple, but alarming – they provide no new insight, but highlight what's important. Global emissions are still going up, whichever greenhouse gas (GHG) we're looking at. China and the United States are the biggest polluters by far, and in the case of China, emissions are still rapidly increasing. Even in developed economies where focus and funding have been somewhat turned towards decreasing carbon emissions, we're still seeing an increase from 2020 to 2021.

Interestingly, per capita view on emissions highlight other countries, where the total amount emitted might not be that high (they may be smaller countries), but they are outstanding when looked at compared to the size of the population. These are mostly oil states, like Qatar, or the United Arab Emirates, the current host of COP28. Not surprising, but worth thinking over.

Apart from the per capita calculation, I did not do an extensive analysis of the data. As a next step, sector by sector results could be compared for similar countries and potential differences or similarities could be highlighted. The full Climate TRACE dataset, that contains details not only on this aggregated level but on actual sites, factories, and mines, is an important source of independent emissions data and further analysis could highlight potential areas of intervention.