Data Report, Stefan Pfahler

May 31, 2024

1 Data Report

This data report contains information on the data sources used for this project, the data pipeline for cleaning and extracting them and limitations of the chosen data.

The main question that I'll try to answer is:

How significant is the impact of the EU Emissions Trading System (EU ETS) in the sectors it sanctions and to what degree does that make an impact on global emissions.

1.1 Data Sources

The data sources I have chosen to answer the main question of this project with, are data of the EU Emissions Trading System in form of the **EU Transaction Log** (EUTL) to find the exact CO2e emissions per sector, the **development of the CO2e price in the EU** to link that to rising/falling CO2e emissions and finally a dataset that contains **global GHG emissions data** to compare the EU with what's happening on a global scale.

Following table contains a quick overview on the data sources, their origin, license, structure and quality. The quality column references the quality dimensions abbreviated in the order “Accuracy”, “Completeness”, “Consistency”, “Timeliness” and “Relevancy”. A ”_” means that the quality dimension is not met.

Data Source	Origin	License	Structure	Quality
EU Transaction Log	The EU	CC BY 4.0 DEED	Structured	AC_TR
Operators in EUTL	The EU	CC BY 4.0 DEED	Structured	ACCTR
CO2e Price Development in EU	Umweltbundesamt	§ 12a EGovG	Structured	ACCTR
Global GHG Emissions	The EU / EDGAR	CC BY-NC-ND 4.0	Structured	ACCTR

The two datasets concerning the EUTL are both licensed under the CC BY 4.0 DEED license, which means the data can be adapted and shared as long as appropriate credit and a link to the license is given. The dataset from the Umweltbundesamt is published under § 12a EGovG, which allows any citizen to freely access and process the data. Additionally when using the data, the Umweltbundesamt has to be mentioned. The global GHG emissions dataset is licensed under the CC BY-NC-ND 4.0 license, which allows to format and redistribute the data as long as the creator is mentioned.

1.2 Data Pipeline

The data pipeline for cleaning the datasets is written in Python using the pandas library. It follows the same pattern for all different data sources that are (1) downloaded from their HTTP resource, (2) cleaned and (3) stored into a SQLite database.

1.2.1 Cleaning and Transformation

As all the datasets were already in a structured state only minor cleaning and transformation steps had to be done. For the EUTL I **joined the two tables** of its .xlsx file into one to improve readability, because the activity types (i.e. sectors) of the companies were stored separate to the main log and were only referenced via an id.

Also the emissions data of the EUTL was split into VERIFIED_EMISSIONS and CH_VERIFIED_EMISSIONS for emissions that stem from the inclusion of Switzerland in the EU ETS. To have a single emissions value to work on, these **two columns had to be added up**. In order to do so I also had to extract the “EXCLUDED” values, which were previously part of VERIFIED_EMISSIONS, to a **new column**.

In the original EUTL dataset, each year was represented by multiple columns that featured e.g. the allocated and verified emissions. To be able to work with the data, I created columns that contained said data for all years by **concatenating each of the previous columns** and **adding a year column** (see table below). I also **removed some columns** that I didn’t count as important for answering my question.

Before:	Country	...	Em_2023	Alloc_2023	CH_Em_2023	...	Em_2008	Alloc_2008
After:	Country	...	Emissions	Allocation	...	...	...	Year

Both the EUTL and the dataset on CO2e prices had a **header that had to be removed**. For the CO2e price dataset, **column names had to be manually added** to the dataframe.

1.2.2 Dealing With Errors

The data pipeline was written specifically for the four static data sources that were selected for this project so it is not accounting for any changing input data. As for errors, none were apparent in the datasets up until now. Errors that might occur during the execution of the pipeline (i.e. network or OS related) are negligible as they are dependent on the system its running on.

1.3 Results and Limitations

The output of the data pipeline is a SQLite database, which contains multiple tables, namely “eu_transaction_log”, “eu_ets_operators”, “co2e_price_development”, “global_ghg_emissions”. All the tables contain structured data that aligns with the data quality dimensions mentioned in the lecture.

I chose SQLite to store the cleaned data, as it was and still is structured data with fixed data types and it also allows for loading the data into a database viewer if needed.

Thoughts on Correctness: The datasets were mostly published by either the EU or the german state. As the data origins are all governmental and not motivated by e.g. company marketing I reckon the data to be accurate. The EUTL in specific is automatically generated by logging all transactions made between accounts in the EU ETS and ensures that these transactions comply with the EU ETS rules. EDGAR on the other hand, which publishes the dataset on global GHG emissions, also claims to provide independent emission estimates through data collected from EU member states or UNFCCC parties.

Thoughs on Completeness: The same governmental argument from before also goes for the completeness of the datasets. I also didn't come across any data I was missing in the datasets.

Thoughs on Representativeness: All datasets grew over the span of decades as they mostly represent the history of CO2e emissions and their pricing over the years up to 2023, which I count as up-to-date. I also reckon I didn't lose or distort any important data during the cleaning process as I didn't add any new data and double checked any data that I removed.

Concluding the data report I have a good feeling for the data analysis and final report. So far the data seems viable and not missing any critical information.