Documentation of the Dataset

**Introduction to the dataset**

The dataset that is used for this research is a dataset that is collected by the United Nations Framework Convention on Climate Change (UNFCCC) and the EU Greenhouse Gas Monitoring Mechanism. This dataset is used to make an index of the individual emissions of greenhouse gasses of the 28 European Union member states between the period of 1985 and 2014. The dataset displays the annual values of the sectors Energy, Industrial Processes and Product Use, Agriculture, Land-Use Change and Forestry and Waste.

The dataset can be found on the website of the European Environment Agency (EEA) through the following link:

<https://www.eea.europa.eu/data-and-maps/data/national-emissions-reported-to-the-unfccc-and-to-the-eu-greenhouse-gas-monitoring-mechanism-14#tab-european-data>.

The file can be downloaded as a zip-file that contains the actual dataset. The used format for the dataset is a CSV file. The file is about 87 MB, which makes it very hard to work with without a rather fast processor. It is possible to use the dataset for personal researches, however it is copyrighted by the EEA. This means that if you want to use the dataset, you have to comply with the rules of the reuse policy of the EEA. More information about this matter is provided on the website.[[1]](#footnote-0)

**Provenance of the data**

As stated before, the dataset is set up by the UNFCCC and the EU Greenhouse Gas Monitoring Mechanism. The information of the dataset is a collection of submissions of the 28 member states of the European Union. The 28 countries have to annually submit files with about 50 tables in which they report the emission values of the different types of industries of that year. To keep these submissions structured and standardized, every country has to conform to a strict set of rules. This is the reason why the dataset is very structured and very clean.

A problem that is specific to the provenance to this dataset is the fact that it is not possible to access the original reports that were handed in by the 28 countries on Eionet. Eionet is a governmental open big data source. In order to access these Eionet files, one needs a username and a password. We were not able to get one, which made us unable to look at the original reports. This could be seen as a problem, because this made us unable to compare the original reports with the information provided in the dataset. Therefore we couldn’t see if the author of the dataset changed certain elements of the original reports before putting it in the final dataset. And it is certain that changes are made because some countries re-submit the reports 2 or 3 times in a year. What they change is not available to the public. Another drawback of the unavailability of these original reports is the fact that we couldn’t see the annotations within the dataset. For instance, if there are no values about the emissions of a certain type of industry every country has to explain the reason why there is no information and submit these in a table. This information can unfortunately only be found in the original reports.

In conclusion it is possible to trace the provenance of the data, but only to a certain point. Because the original data is not authorized for the average user of Eionet, it is not clear what has been done to the data before it entered the dataset. Especially if there are several re-submissions annually. For this reason we have to assume that the given information per country is correct, however, we can not assume that the reported data is factual. So the data is public in name only but is still private in practice, which makes curating the dataset nearly impossible.

**Curation of the dataset**

***Description of the plan***

The dataset itself consists of about 600332 rows of information. This means that processing the data would take a pretty long time because of the size of the dataset. For that reason, we chose to make a new dataset that only includes the data we want to use for the research.

We wanted to filter out two different things out of the dataset:

* the five different sector names that have something to do with transport (Domestic Aviation, Road Transportation, Railways, Domestic Navigation and Other Transportation). By filtering these five things out of the dataset, you get only the relevant information because we focus on the transportation part within the dataset.
* Four different years (1990, 2000, 2010, 2016). We chose these four years because they show the beginning, two years in the middle and the end of the dataset. With the data of the four different years, we can determine the emissions greenhouse gasses per transportation method, and with that the its share in the total emission of the transportation sector. This might give us a good overview of the growth per transportation method within the transportation sector.

***Method***

First we loaded the dataset in OpenRefine. The dataset is a .CSV file, which made it easy to implement the dataset in Openrefine. In OpenRefine, wemade a text facet of the Sector\_name to filter out the sectors that have something to do with transportation. We did this by excluding the numbers 1.A.3.a - 1.A.3.e, which are the five different transportation methods mentioned in the dataset. With the filtering, the dataset has gotten a lot smaller; from 600332 row to 19200. we exported the filtered dataset into a new .CSV file to make further investigations, because the processing of the smaller dataset will be a lot faster.

We opened the new .CSV file again in OpenRefine to filter out the years we want to include. We did this by making a new text facet of the years. With the text facet, we included the 4 years that we want to investigate for the research (1990, 2000, 2010 and 2016).

While working with in OpenRefine, we found out that the UNFCCC also wants the countries to count the amount of emissions for every type of pollutant (CO2, CH4 or N2O). However, we are not interested in one specific type of pollutant, so would mean that we have to add the three different types of pollutants up to each other to find out the total emissions per transportation method. Luckily, the editors of the dataset already added something similar; they already made a variable that is called ‘All greenhouse gases - (CO2 equivalent)’ which are all the pollutants added up to each other. We decided to also include this variable and filter out the reset, because it gives a quick overview of the total amount of all types of greenhouse gases per transportation method.

In total, we applied 3 filters on the dataset. We filtered on the Sector\_name for the transportation method, on the Year to filter out the relevant years and on the Pollutant\_name to find out the total amount of emissions of greenhouse gasses per transportation method. With all the filtering done, the dataset now only contains 700 rows with relevant information for our research. The 700 rows should be correct, because the filtered dataset includes 35 countries, that report 5 different transportation methods over 4 different years. 35\*5\*4 = 700

However, while working with the dataset, we realized that the dataset consists of 35 different countries, while the European Union only consists of 28 different countries. So we made a text facet again, and we found out that besides the 28 countries of the EU, another 5 countries are included in the dataset. These 5 countries are Iceland, Liechtenstein, Norway, Switzerland and Turkey. Besides that, the editors also added ‘EU28’ and ‘EU (KP)’ as countries. EU28 is the sum of the 28 European Union members and ‘EU KP’ is probably the sum of the countries who participate in Kimberly Process (KP). We decided to filter EU28 and EU (KP) out of the dataset, because we want to do our own calculations with the dataset and do not want to rely on ready made calculations. With these two variables filtered out, the dataset now has 660 rows. This is our final filtered dataset.

***Making graphs with Excel***

In order to find out which subsector of the transportation sector has the biggest growth, we wanted to make pie charts of the four different years and a graph that shows the percentage growth per subsector. We wanted to do this in Microsoft Excel, since this is a very effective and easy platform for making graphs and adding up the values of the subsector.

The first step was to convert the CSV file to a Excel file with OpenRefine. Once we had done this, we deleted all the unnecessary data of the dataset in order to make the graphs. The only data we kept were the countries, sector\_name and amount\_of\_emissions. This enabled us to add all the values of the emissions of all the countries up to each other, to find out the total emissions per subsector per year. Once we had got all the emissions per subsector, we were able to find out the total emissions of the transport sector per year by adding up all the emissions of the subsectors.

After the preparations, we proceeded by making pie charts with the new found information. We chose pie charts, because we thought that pie charts give a good overview of the share of each subsector in the total transport sector. We used the built in function to make pie charts in Excel.

Then we decided to focus solely on the aviation industry and created a line graph by calculating the amount of emissions for 1990, 2000, 2010 and 2016 respectively. This graph we used to support our conclusion.

1. <https://www.eea.europa.eu/legal/copyright> [↑](#footnote-ref-0)