Milestone 2: Data Collection Pipeline

Bayrakceken, Kudret Aras 03669629

Belkhiria, Zied 03653792

Bueno Ulacia, Ion 03726897

Egger, Maximilian 03735004

Kern, Max-Emanuel 03673151

Krüger, Philipp 03673587

Martín Cruz, Daniel 03727385

Tarasewicz, Damian 03734755

June 28, 2020

Contents

1. Introduction

Our team decided on tackling research question No. 2: *To what extent will the Covid-19 pandemic contribute towards reaching goals stated in international agreements*. During our research on this question, we concluded that it is best to focus on the *Paris Agreement*, as it gives a greater frame also to other international agreements. The *Paris Agreement* got a lot of publicity when it got ratified by most countries of the world. It thus resembles the first global, legally binding agreement to fight global warming. When the US under president Trump decided to leave the agreement and climate activists started the *Fridays for Future* movement, the problem of global warming got more attention worldwide. However, as the COVID-19 pandemic began, more urgent problems arose and global warming only came in second. We now want to investigate how the ongoing pandemic affected global greenhouse gas emission and see how this could actually help fulfilling the ambitious goals of the *Paris Agreement*.

We searched through several fields to find recent and reliable data that we could use to investigate our question. These fields include not only COVID-19 and global greenhouse gas emission but also mobility, weather and of course international agreements. We gathered reliable data, established a data processing pipeline and visualized the data in several plots. Furthermore, we created an interactive website that displays our data. To access this website, please use the following link `https://ami-group1-dashboard.herokuapp.com`. The downloaded datasets can be found in our Git repository within the subfolder */datasets/*.

2. COVID-19 Datasets

The main task in our research is to investigate the impact of the Covid-19 pandemic on greenhouse emissions, and its contribution to the countries, in achieving their goals of reducing emissions. So it was necessary to gather data on Covid-19 statistics of the countries, where it is available. These statistics include, but are not limited to, the number of cases, deaths, recoveries, and ratios of these numbers to the population. We believe these numbers are responsible, together with each countries response to the pandemic, for the change in greenhouse emissions during the pandemic. This data is vital to deduce the correlations between other data we have and predict the future of greenhouse emissions. In general, the datasets we have found and implemented are

- easy to access online or offline.
- updated frequently (at least daily).
- containing both Covid-19 data and coordinates of the corresponding region.
- containing data for both countries and smaller administrative regions.
- not containing general population data.

2.1. Data Sources

We have implemented ways of accessing data for two distinct datasets since it is best to compare different data sources for accuracy. Since external population data was needed for both datasets, a dataset [1] was downloaded and preprocessed to contain only the most recent population info, which is from 2018 except for Eritrea (2011).

2.1.1. Bing

The first dataset is provided by the Bing search engine [2]. It is a CSV file, hosted on GitHub and updated every day by Microsoft. They collect data from multiple reliable sources, including WHO, state health departments, Wikipedia, etc. Since it is hosted as a single file on GitHub, accessing the data in real-time is pretty straight-forward, by just providing the link to the file. The statistics in this dataset are the daily number of cases, deaths, recoveries. Moreover each there are samples for administrative regions within the countries and additional features are provided with each sample; last update date, coordinates, the country name in multiple ISO formats. The downside of this dataset is that it does not include any population data, so to calculate the ratios of cases, deaths, and recoveries to the general population, the population for each country had to be accessed from a different dataset [1].

Firstly smaller administrative regions than countries were cleaned from the dataset since we are looking at the emission goals of countries and processing data for smaller administrative regions would be costly. Then, to inspect the temporal change of the statistics for a selected country, the database can easily be filtered by the name of the country (since the samples are already sorted by date). But to inspect and create graphs on the most actual data, all the samples but the most recent ones for each country were deleted. From there using the most actual data, all the features we have are drawn on world maps, with the help of Cartopy[3].

2.1.2. Johns Hopkins CSSE

The second dataset we use is provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University and gathered from multiple sources [4]. In our project we access the dataset through https://covid19api.com/, which provides easy access to the dataset, using an API request. Just like the Bing dataset, it includes the number of cases, deaths, recoveries, alongside the coordinates of the corresponding country or administrative region. This makes it easier to compare them but population data for each country must be taken from another source to calculate the ratio of the statistics to the general population. According to covid19.com, the dataset is updated multiple times a day. This can be an advantage compared to the dataset of Bing, which is updated daily.

2.2. Processing data

- Importing data either online or offline.
- Choosing only country-level administrative regions.
- Selecting either most recent sample of each country or multiple samples over time for a single country.
- Visualizing either recent statistics of all countries or temporal data for a single country.

The world maps are drawn with Cartopy [3]. Other figures are drawn with Matplotlib [6].

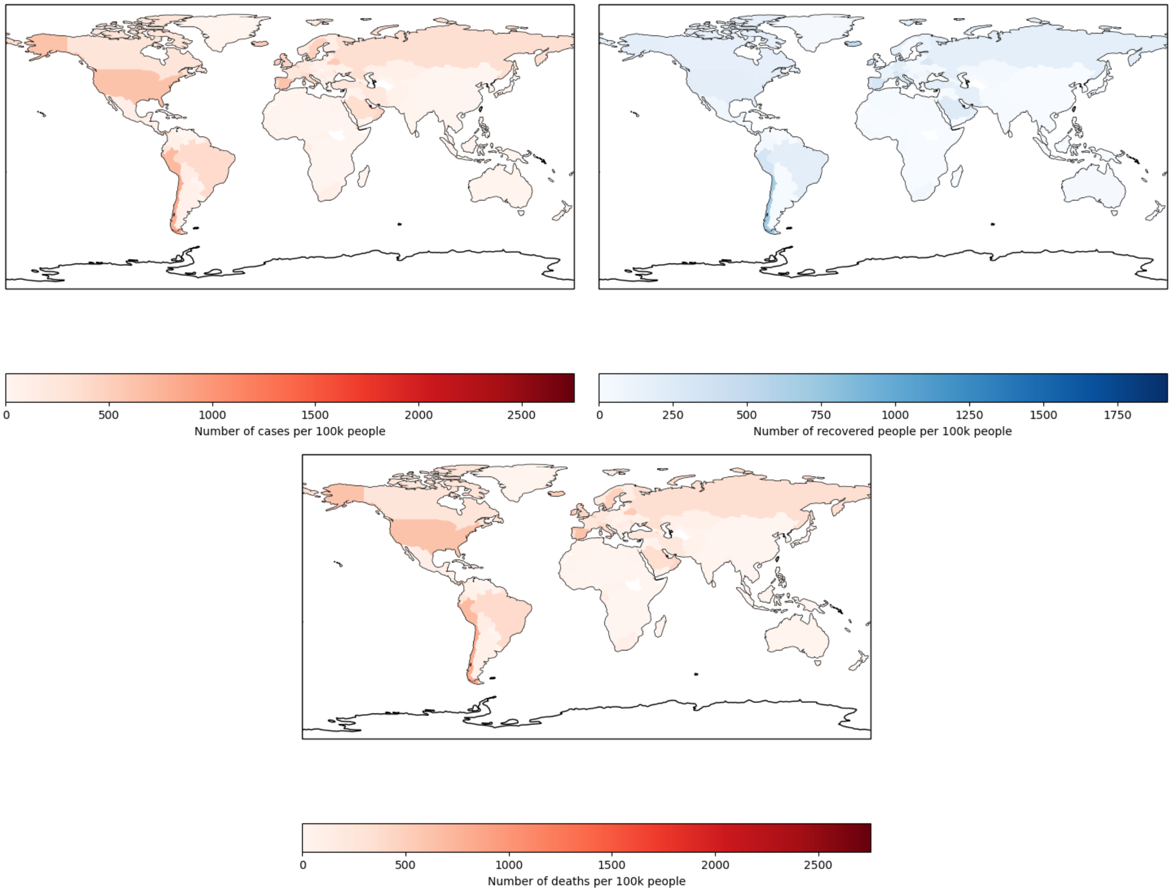


Figure 1: Covid-19 rates for each country.

2.3. Conclusion

Having detailed Covid-19 data is a necessity for our research, since the effect of the pandemic on greenhouse gas emissions will be researched. In the end the datasets we have implemented are frequently updated, easy to access and use, consistent with each other, and have detailed information on cases. Therefore, we don't prefer one dataset over the other. Since we can access temporal data from individual countried throughout the pandemic, it will be easier to model the correlation between Covid-19 cases and greenhouse gas emissions.

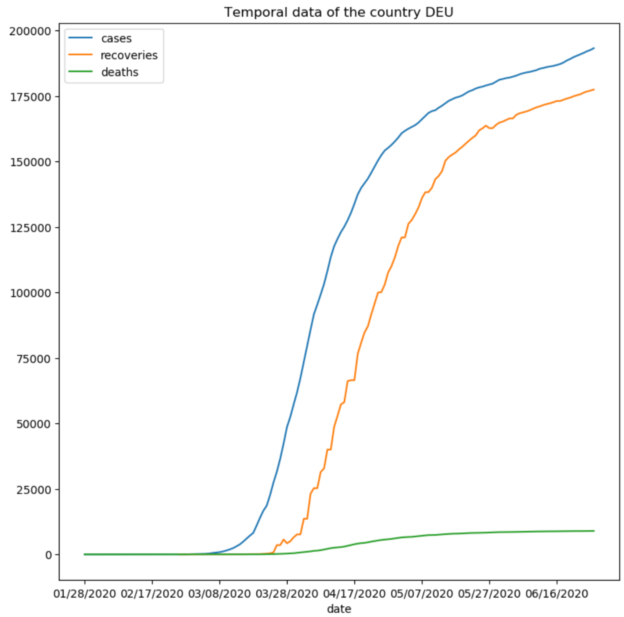


Figure 2: Temporal data of Germany. Drawn by using Bing dataset.

	Country	CountryCode	Province	Confirmed	Deaths	Recovered	Active	Date
0	Germany	DE		8	0	0	8	2020-02-01 00:00:00+00:00
1	Germany	DE		10	0	0	10	2020-02-02 00:00:00+00:00
2	Germany	DE		12	0	0	12	2020-02-03 00:00:00+00:00
3	Germany	DE		12	0	0	12	2020-02-04 00:00:00+00:00
4	Germany	DE		12	0	0	12	2020-02-05 00:00:00+00:00
...	...	...	...	...	...	...	...	...
126	Spain	ES		241310	27135	150376	63799	2020-06-06 00:00:00+00:00
127	Spain	ES		241550	27136	150376	64038	2020-06-07 00:00:00+00:00
128	Spain	ES		241717	27136	150376	64205	2020-06-08 00:00:00+00:00
129	Spain	ES		241966	27136	150376	64454	2020-06-09 00:00:00+00:00
130	Spain	ES		242280	27136	150376	64768	2020-06-10 00:00:00+00:00

	Country	CountryCode	Province	Confirmed	Deaths	Recovered	Active	Date
0	China	CN	Zhejiang	599	0	21	578	2020-02-01 00:00:00+00:00
1	China	CN	Ningxia	26	0	0	26	2020-02-01 00:00:00+00:00
2	China	CN	Qinghai	9	0	0	9	2020-02-01 00:00:00+00:00
3	China	CN	Shaanxi	101	0	0	101	2020-02-01 00:00:00+00:00
4	China	CN	Shandong	206	0	3	203	2020-02-01 00:00:00+00:00
...	...	...	...	...	...	...	...	...
4318	China	CN	Shanxi	198	0	198	0	2020-06-10 00:00:00+00:00
4319	China	CN	Sichuan	582	3	560	19	2020-06-10 00:00:00+00:00
4320	China	CN	Tianjin	195	3	189	3	2020-06-10 00:00:00+00:00
4321	China	CN	Tibet	1	0	1	0	2020-06-10 00:00:00+00:00
4322	China	CN	Xinjiang	76	3	73	0	2020-06-10 00:00:00+00:00

Figure 3: Example of samples from the Johns Hopkins dataset.

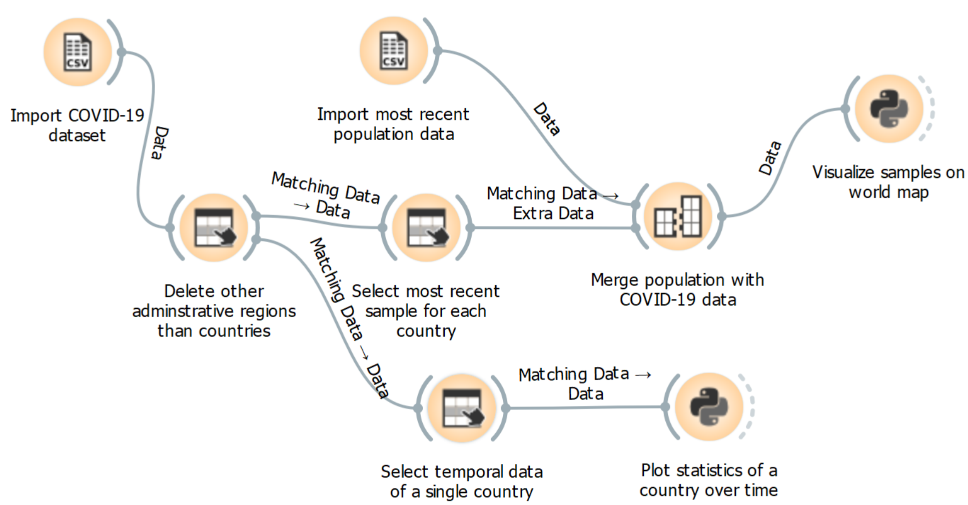


Figure 4: Our general pipeline of processing Covid-19 datasets. Drawn using Orange [7]

3. Greenhouse Gases Datasets

If we want to know what impact COVID-19 could have towards reaching goals stated in international agreements respect climate change, it is critical to have a sufficient amount of current and previous data in greenhouse gas emissions.

We searched for datasets of greenhouse gas emissions and these are the most relevant sources we found:

- At first order we thought **NASA** APIs [8] could be useful. There are many datasets about emissions with great detail. The negative points was accessing to them, being really difficult and, most importantly, there was no data available within a national scope, which is important to compare it with the goals stated in the international agreements.

- From the **Global Monitoring Laboratory** [9], we get global datasets of the main four greenhouse gases: CO_2, CH_4, N_2O and SF_6 which are critical in the climate change. With respect to this source we get an important amount of samples which includes current data. We can distinguish three datasets respect the time consideration: monthly, annual and annual increase. Appart from that, there is a weekly updated CO_2 dataset. The main drawback is that the data is not classified by country or by sector.

- To deal with the lack of data by country and by sector, we got these datasets from the **Emissions Database for Global Atmospheric Research** [10]. The only problem is that these datasets are not updated, until 2018 in the case of CO_2 and until 2015 when it comes to CH_4 and N_2O, being all samples by year. Apart from that, there is no data about SF_6.

- We have found a complete and popular dataset from the **International Energy Agency** [11] with data which could be very useful. The main problem is to access it, since we have to pay an expensive license.

- There are complete and current datasets from cities as **Madrid (Spain)** [12] or parts of determined countries as **India** [13]. Due to data availability of other regions, it could very difficult to get a complete amount of data if we want a broader scope as Europe.

- An interesting source is the **Air Quality Historical Data Platform** [14], where we can get current data of stations in determined cities. The negative point is the monitorized gases are not the main cause of the climate change.

There are differences in the collected data which led us in not selecting all sources. Time range is not the same, we have current data or until certain year as 2018. In addition to that, data is collected with different frequency, from annual to weekly in some cases. The spatial scope varies from a global perspective to by countries or cities. It is important to have into consideration the data accessibility, since in some cases, the access requires payment or the APIs request is difficult to manage.

Taking into account all the pro's and con's of each source, we decided to use the ones from the **Global Monitoring Laboratory** [9] due to the current data and the frequency collection (monthly). Besides that, mention it provides us a global perspective. We also used the **Emissions Database for Global Atmospheric Research** [10]. The main reason was it has annual data classified by country and by sector in the case of CO_2.

In order to prepare the datasets for their final use, we followed different steps to get them cleaned, loaded in our python code and visualized so that all the members can quickly check the obtained data regarding gas emissions.

For the datasets from the Global Monitoring Laboratory we first read the corresponding csv or txt file using only the columns we want which are date (that can be year, year and month or year month and day) and mean. For the case of monthly CO_2, instead of the "mean" column, we selected the "interpolated" one because they handled missing data calculating these values using the trend that the CO_2 emissions were following and the data from other years in the same season. In other cases where we had missing values and the interpolation option does not exist, we decided to fill them with the previous value in the time series. After that, we put together all the datasets that had the same temporal scope (yearly, monthly, weekly and annual increase).

When it comes the EDGAR datasets, we end up generating four different dataframes which are: CO_2 by country, CO_2 by country and sector, CH_4 by country and N_2O by country. We obtained these dataframes from 3 different excel files, one for each gas, through coverting them to csv files and reading them in a proper way using pandas. Another step useful for plotting and interpretation, was adding one more variable to the list adding up the contributions of all the 28 countries of the European Union. However, we also kept the data of all the 28 countries because each one will have individual COVID case numbers. We did this because in the international agreements related to climate change, the whole European Union is represented as one actor and having the progression of all the Union as a whole can help us with the model and its interpretation.

Finally, for all these dataframes, we generated different pdf files with plots to visualize all this information just by taking a quick look. For those dataframes where we have data by country, we plotted the data of the eight main countries which contribute to the emission of greenhouse gases.

In order to have a global perspective of emissions, we collected annual data of CO_2, CH_4, N_2O and SF_6. Also, the annual increase of these gases [9]. Important to mention each gas has its own units (see the legend), since plotting them in the same scale gives us a non intuitive graph.

As it can be seen, we collected more samples for CO_2, due to the availability of data.

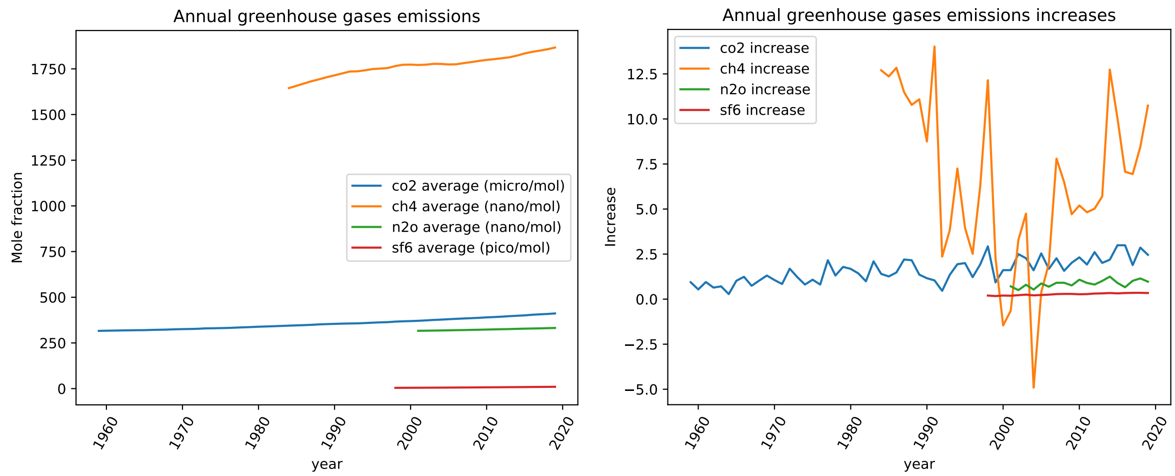


Figure 5: Annual global emissions of greenhouse gases.

If we want more detail, we can take a look to monthly data or even weekly in the case of CO_2 [9].

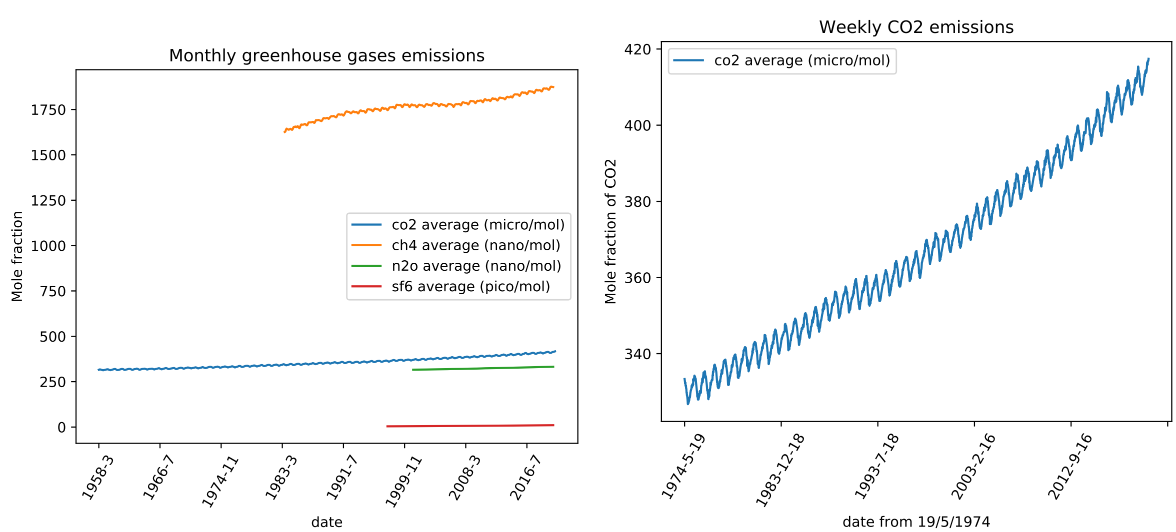


Figure 6: Monthly and weekly global emissions of greenhouse gases.

Depending on the scope we select to study, it is going to be useful to have detailed information, as emissions by country or by sector.

In the case of CO_2, we collected emissions by country and by sector. The eight big contributors to greenhouse gases are plotted. In addition to that, five sectors are studied: power industry, buildings, transport, other industrial combustion and other sectors [10].

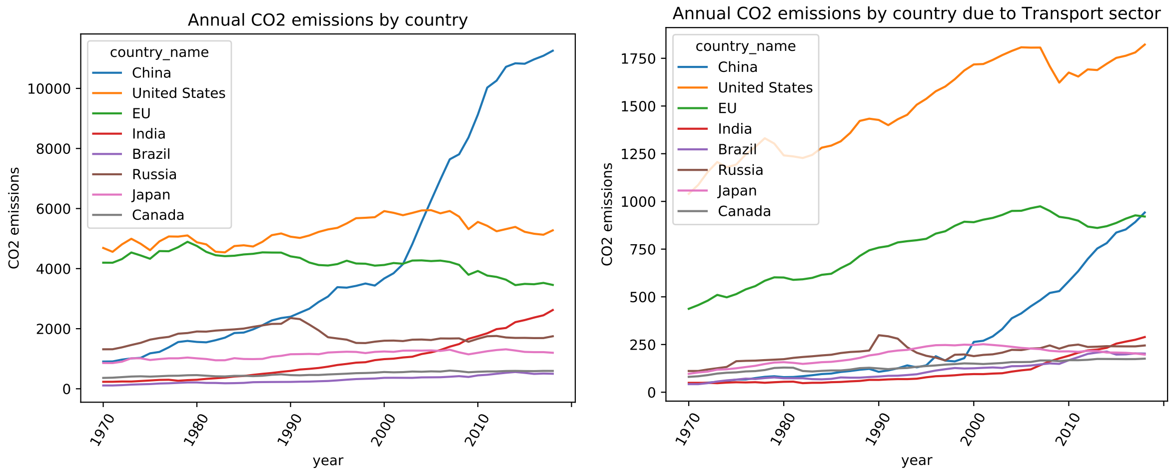


Figure 7: CO_2 emissions by country and by sector.

For CH_4 and N_2O we also have the emissions by country [10].

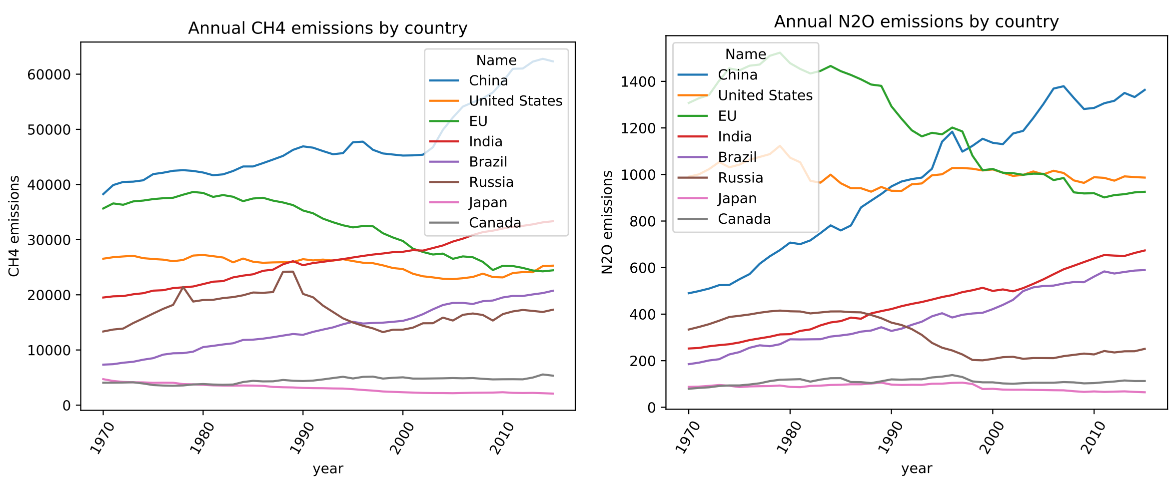


Figure 8: CH_4 and N_2O emissions by country.

We can state that the data collected and prepared from the two previously indicated sources, **Global Monitoring Laboratory** [9] and **Emissions Database for Global Atmospheric Research** [10], will be useful. The reasons are the range of detail respect time and space. In addition to that, we have information regarding sectors and determined gases, which can be critical if we are looking for a certain problem.

Obtained greenhouse gases emissions data is probably the most relevant information for the project together with COVID-19. For this reason it was very important for us to get good datasets in order to achieve a proper model. Getting data of emissions by country allows us to predict the emissions in the following years to compare that with the commitments stated in international agreements. Besides that, having data with respect to determined sectors is going to be useful relating the decrease of activity and greenhouse emissions. The only drawback of these datasets is that they are not updated up to 2020. Latest samples are from 2018 in the case of CO_2 and 2015 in the case of CH_4 and N_2O. To compensate this lack of up-to-date information, we included some global datasets provided by the Global Monitoring Laboratory updated up to May 2020 and with a deeper detail in time (monthly, weekly). Combining these two sources of information, the need of greenhouse gases emissions data is satisfied.

4. Mobility Datasets

In order to get insights about the possible impact of changed mobility behavior on greenhouse gas emissions, it is important to use different kind of mobility data. The goal by using those datasets is to find a correlation, if there is one, between the mobility behavior and the greenhouse gas emissions. Of course, other aspects that affect the atmospheric greenhouse gas levels also have to be taken into consideration.

Mobility behavior relates to various aspects in our life. That's why it is important to have many different kinds of datasets for this topic. Apple and Google recently published mobility reports for almost all countries in the world. Both datasets might be useful in our case due to the fact that they are differently organized. Besides the mobility reports, aviation data might also be interesting to analyze. Since aviation is quite representative for the long-distance travel behavior of the people, this might give some important insights into long-distance traveling compared to regional mobility aspects provided by the mobility reports. To ensure the correctness of the collected aviation data, two different datasets will be used and compared to each other. In the following, the different datasets and the information they contain is explained in a more detailed way.

- **Apple Mobility Trends:**
  The database of Apple's mobility trend report is updated daily and reflects the amount of requests for directions per day in Apple's cards app.[15] The data is normalized with a baseline on January 13th, 2020, and shows the further development in percent in respect to this baseline. In total the datasets is split up into data for driving, walking and transit data. This could be extremely useful in our case, since the different kinds of mobility data probably have a different impact on greenhouse gas emissions.
- **Google Mobility Report:**
  The Google mobility reports are structured in a different way. Here, information are given for movement patterns at different categories of geographic places.[16] The dataset contains information about movement behavior in transit stations, workplaces, parks or retail and recreation. Together with the relationships between different movement patterns and their impact on the atmospheric greenhouse gas level, we will try to find insightful correlations. Determining such correlations, we could predict the impact of an increasing movement behavior after the COVID-19 crises and its impact on the greenhouse gas emissions.
- **OpenSky Network Flightlists:**
  The OpenSky Network provides air traffic data for research topics. The specific dataset we use was updated every month from January 2020 up to May 2020 and hopefully also continues to be updated in the future. It includes data for all flights within this period, including the flight numbers, origin and destination airports and timestamps. The advantage of this dataset is that it contains very detailed data. For example, we could determine the number of flights per day which took off or landed at a specific location. Also, we could determine the number of total flights per day. In this case, we decided to analyze the overall flight hours per day because this is most directly correlated with the greenhouse gas emissions. The disadvantage by using this dataset is that it contains lots of data. Therefore, the pre-processing steps take a relatively large amount of computational resources. The data sources can be found by following this reference: [17].
- **Flightradar24 Tracking Statistics:**
  Flightradar24 provides historic and live air tracking data. The database is separated into statistics for all flights and only for commercial flights. In each of the cases, the number of flights per day is given as well as a seven-day moving average. This dataset provides helpful insights into the air traffic by day since January 27th, 2020 in a very clear way. The disadvantage of this source is that it does not provide any detailed information about the length of the flights. For example, drone flights are tracked by Flightradar24 which do not affect greenhouse gas emissions. That's why this dataset can only be used in a limited way for our problem, but still it provides some good information about movement patterns in general during the COVID-19 crisis and in the future. In addition, it has to be taken into account that aviation in general affects the overall greenhouse gas emissions only in a small way. The dataset can be found here: [18].

For all the collected datasets different pre-processing steps had to be taken. In order to get an intuition about the collected data, the goal was to also create plots which present the data in a comprehensible way. All of the pre-processing steps are bundled in a loader class.

The pre-processing of the Apple data mainly consists of specifying the information to extract from the dataset. Therefore it has to be chosen from the three different transport types driving, walking and transit. Additionally, the countries to be extracted have to be specified. If data for the requested country is not available, this country will be ignored in further processing steps. By use of this structure, we are then able to dynamically extract the kind of information we need depending on our future research focus. After extracting the relevant information from the csv file and dropping the irrelevant ones, the data is restructured as a pivot table in order to separate the data by countries and index the samples by date. Missing values are filled with the last valid sample. An exemplary plot for transit data in four European countries is shown in Figure 9.

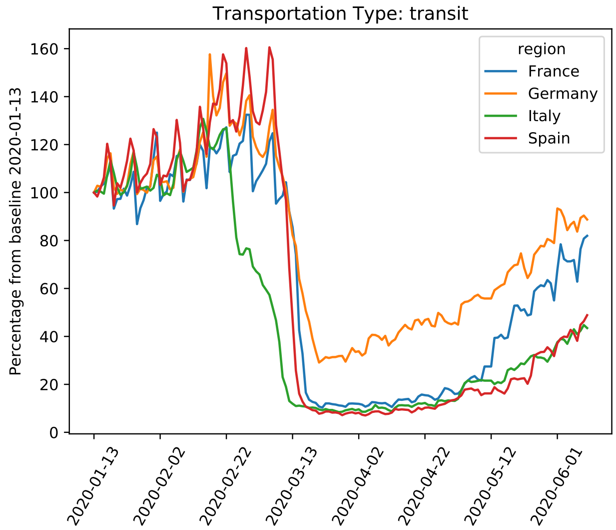


Figure 9: Exemplary plot for Apple mobility trend data.

The pre-processing of the Google mobility reports basically follows the same steps. Also the method for handling missing values is the same. In contrast to the Apple datasets, other information types have to be specified here as discussed in the listing above. A demo plot for the movement patterns at workplaces in different European countries is shown in Figure 10.

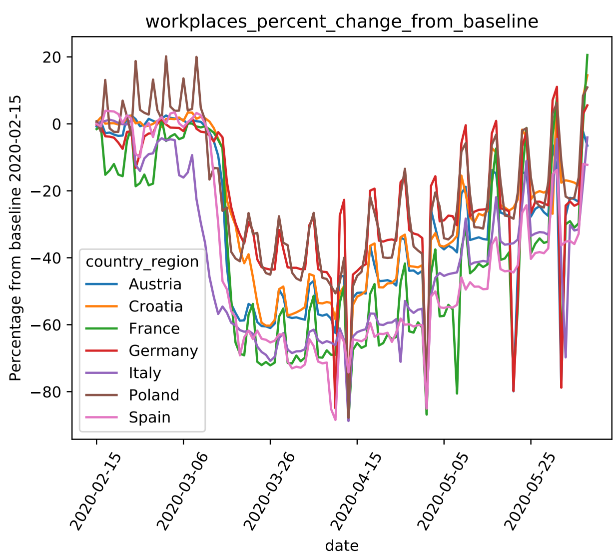


Figure 10: Exemplary plot for Google mobility report data.

Regarding the OpenSky dataset, the pre-processing was computationally more expensive. Since the data were provided in monthly packages, they had to be merged at first. In order to save computational resources, only necessary parts of the dataset were extracted. Since we wanted to extract the overall flight ours per day, the first seen and last seen timestamp was used to calculate the duration of every flight listed in the dataset. Afterwards, a pivot table is created which sums up the flight hours for each day and uses the days as indices. Finally, the day timestamp is modified to only show relevant data.

Since these steps take a while when running them on a desktop computer, the results were stored in a new csv file containing the overall flight hours per day since January 1st, 2020. The result can be seen in Figure 11.

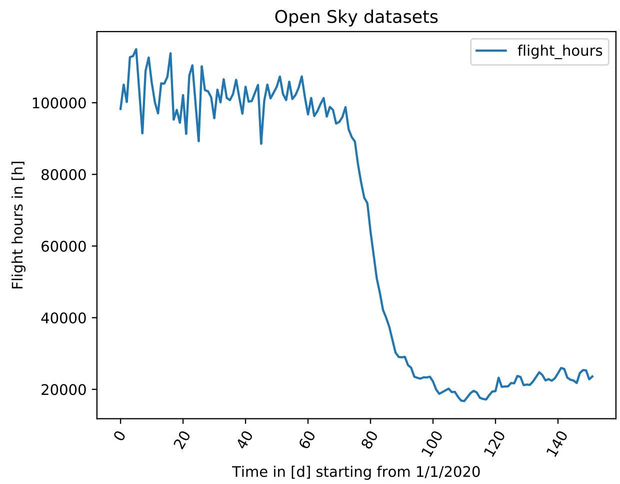


Figure 11: Flight hours per day extracted from the OpenSky datasets.

The Flightradar24 dataset was already well-prepared. It simply had to be merged into one data frame. The content of this dataset is summarized in Figure 12.

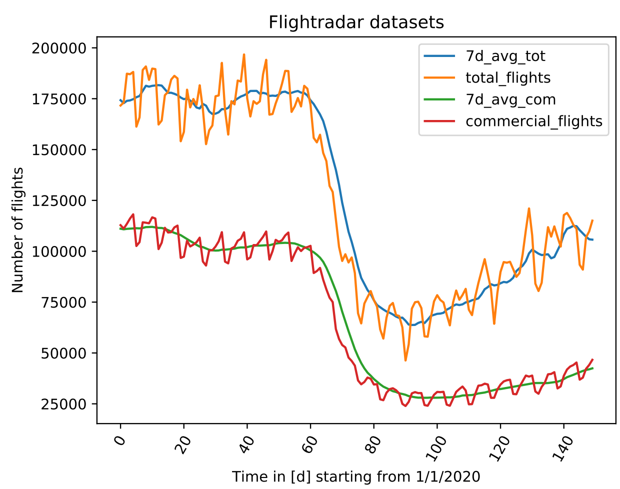


Figure 12: Number of commercial and total flights provided by Fightradar24.

All of the collected data might be useful for predicting the influence of COVID-19 on global greenhouse gas emission goals. Especially for predicting the impact in different countries, the mobility report data sets by Apple and Google could be very helpful since they differentiate between movement patterns in different countries. When it comes to transportation, in particular the Apple data set could be very useful because it is separated into different kinds of traveling. This could be extremely useful as different kinds of mobility also have different impacts on greenhouse gas emissions. Together with some side information like the emissions of vehicles, we could try to find correlations between the greenhouse gas levels.

The air traffic data by Flightradar24 has to be reviewed critically. That's because the number of flights does not contain any information about the duration of the flights. Still, it provides an overall intuition about the air traffic and in particular supports the information provided by OpenSky. These datasets could be even more useful since it directly presents how long air planes were flying. Together with the information about average fuel consumption and air plane emissions we can probably use this data to predict the impact of COVID-19 on international climate goals. On the other hand, it has to be taken into consideration that direct emissions from aviation only contribute to about two percent of global emissions [19].

Depending on the machine learning models we are going to use, there will still be some more pre-processing steps required to finally use the datasets for training. One possible strategy to combine the obtained datasets could be to directly take into account the impact of the various contributors to greenhouse gas emissions in percent. This way the information could be combined to one data set. Since we obtained data for various greenhouse gases, this step could also be done for specific greenhouse gases independently in order to improve the accuracy of our predictions. All of the described datasets provide information for roughly the same time interval. The smallest amount of time is covered by the Google data set. That's why an approach for combining the datasets could be to start in February 2020. As it can be seen by inspecting the other datasets which provide information since January 2020, until February the changes in mobility were almost negligible. That's why it will be enough to consider data since February 2020.

5. International Agreements Datasets

To investigate to what extent the COVID-19 pandemic will help countries to reach the goals of the international agreements such as the *Paris Agreement under the United Nations Framework Convention on Climate Change*, we needed to find and understand the goals of such agreements. Specifically, we took a closer look at the Paris Agreement and the agreements of the EU. When we researched the agreements of the EU members, we found out that these agreements are incorporated in the Paris agreement. Thus, we focused on gathering data about the Paris Agreement and handle two international agreements at a time.

5.1. How Does the Paris Agreement Work?

Basically, the *Paris Agreement* states that every country ratifying the treaty acknowledges anthropogenic climate change, commits to the well below 2 °C and further aims to limit global warming to 1.5 °C. To reach these goals, every country has to define its contribution to fight climate change by itself. These so-called **n**ationally **d**etermined **c**ontributions (NDCs) have to be submitted every five years with new, even stricter goals. Furthermore, countries are also allowed to define the same goals in a group, as for example the member of the EU did.

5.2. Gathering and Pre-processing Data

The great challenge in finding data about each country's goals, is that every country defines their own goals, without a greater frame. Thus, countries can for example define a cut in greenhouse gas emissions given as a percentage referring to the emissions in a certain year. It is also possible to state a total amount of greenhouse gases a country wants to reduce its emissions. Also, in some NDCs, countries state that they will build more climate-friendly energy plants. Other, mainly less industrialized countries had a mix of unconditioned and conditioned goals. From these few examples, we can already deduce that the data we can gather from the NDCs is highly non-uniform.

Additionally, there is no readily available data about the NDCs. This means, our team would have to read through every NDC and extract the relevant information by hand, consuming an incredible amount of manpower. Luckily, we found a website[1], in which all NDCs are summarized to the very core of information. From this, we started to extract the designated reduction in greenhouse gas emissions for several countries. However, we quickly realized that this would still consume too much man power. Therefore, we decided to only take a closer look at the eight largest contributors, covering more than 70% of the worlds greenhouse gas emissions in 2017.[2] These contributors are China, the US, the EU, India, Brazil, Russia, Japan and Canada in descending order.

Most of the contributors we considered stated a range of emission reduction they aim for. To compare the data more easily, we simply took the mean of minimum and maximum percentage of emission reduction in these cases. Apart from that, the data is rather uniform and we thus did not have to pre-process it any more. To the data we had from the website[3], we added the share of CO_2 emissions in 2012 and 2017 and the total CO_2 emissions in the referenced year.

5.3. Data and Plots

Country	Share of CO_2 Emissions [%]	Reduction Goal [%]	Compared to Year	Fulfill by Year
China	24.48	*62.5	2005	2030
USA	13.63	*27	2005	2025
EU	10.94	40	1990	2030
India	7.18	*34	2005	2030
Russia	4.48	*27.5	2005	2025
Japan	3.35	26	1990	2030
Canada	1.76	30	2013	2030
Brazil	1.68	37	2005	2030

Table 1: Countries are in descending order by their respective share in global CO_2 emissions in 2017. Greenhouse gas emission reduction goals with an asterisk * are mean values of the range of maximum and minimum reduction goal of the respective country. Further, the year every country refers its reduction goal to and the year by that each country wants to reach its goal is given.

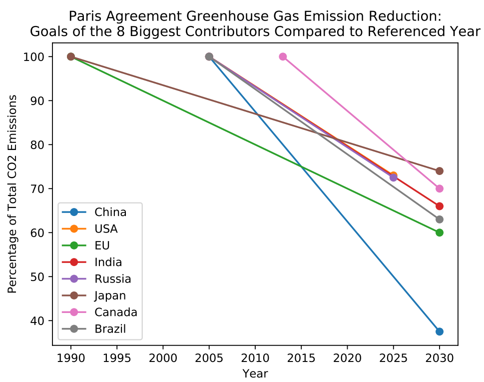
[1] See `https://www.carbonbrief.org/paris-2015-tracking-country-climate-pledges`

[2] See `https://www.worldometers.info/co2-emissions/`

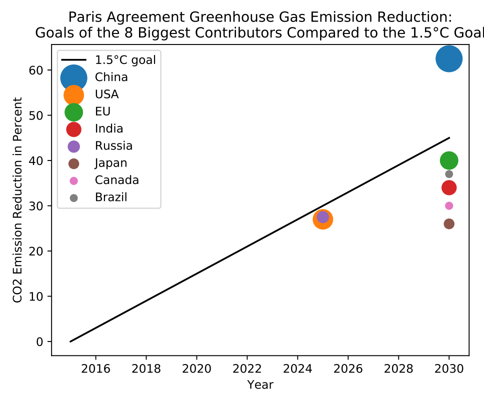
[3] See `https://www.carbonbrief.org/paris-2015-tracking-country-climate-pledges`

In Table 1 we summarized the most important part of the data we gathered for the greenhouse gas emission reduction goals. With this data, we created the plots in Figure 13a and Figure 13b. In Figure 13a we simply draw a line from the starting point to the goal of each country. The starting point is at the year each country refers to in its goals, the value is thus set to $100\,\%$ for every respective country. The end point is at the year every country wants to achieve its goals at the height $100\,\%$ minus the reduction goal. The height therefore corresponds to the greenhouse gas emission level each country wants to achieve in the respective year, compared to the year it referred to. From the slopes, we can deduce the rate of greenhouse gas emission reduction in %/year. This can be used as a measure of how ambitious each country is.

To put the ambitions of each country into perspective, we also need to consider its contribution to the global greenhouse gas emissions. We visualized this in Figure 13b, a bubble plot where the size of each bubble represents the country's share of greenhouse gas emissions. The black line represents the lower boundary of emission reductions to reach the $1.5\,°C$ goal[4]. As we see, only China's goals are ambitious enough to actually reach the $1.5\,°C$ goal all countries aim for. Still, the US, the EU and Russia ratified goals that are at least close to the $1.5\,°C$ goal. Other countries are less ambitious.



(a)



(b)

Figure 13: **a.** Line plot for each country's goal compared to the reference year. The y-axis represents the greenhouse gas emissions of each country in percent, the x-axis the year. Each line starts at (referred year, 1) and ends at ($1-$goal, targeted year). The slopes thus represent the rate of reduction per year in %/year. This plot does not state anything about absolute data or the contribution of the countries to global greenhouse gas emissions however. **b.** Bubble plot that visualizes how ambitious each country is. The y-axis represents the greenhouse gas emission reduction goal in percent, the x-axis the year. The size of each bubble corresponds to its share in global greenhouse gas emissions. The black line gives the necessary emission reductions to reach the $1.5\,°C$ goal.

5.4. Usage of Emission Reduction Goals Data

The data we gathered is crucial for our project. With the data, we can compare the greenhouse gas reduction due to the COVID-19 pandemic with the actual goals each country set for itself. From that, we can quantify how much this reduction contributes to the goals. As we have data of the emitters accounting for over two thirds of global greenhouse gas emissions, we not only see how COVID-19 related emission reductions help some individual countries to reach their goals but also see its global impact. For this reason, the data about emission reduction goals – along with COVID-19 and greenhouse gas emission data – counts to the most important data sets of our project.

[4]We need to cut emissions by 45% by 2030, see `https://www.scientificamerican.com/article/global-promises-to-reduce-co2-are-falling-short-of-1-5-degree-c-warming-goal/`

6. Conclusion and Outlook

In summary, we have gathered data concerning four big aspects that are important for answering to what extent the Covid-19 pandemic will contribute towards reaching the goals stated in the Paris Agreements. Each dataset was found to be crucial for solving the task. To conclude this overview, we list our main decisions and findings:

COVID-19:

- Frequently updated datasets with high temporal resolution per country.
- Preferred datasets: **Bing**, **Johns Hopkins CSSE**.

Greenhouse Gases:

- The datasets discriminate by type of gas, spatially, temporal and by industry sector with reasonable resolution.
- Preferred datasets: **Global Monitoring Laboratory**, **Emissions Database for Global Atmospheric Research**.

Mobility:

- By using the datasets, correlations of regional and global transport with greenhouse gas measurements are accessible.
- Preferred datasets: **Apple Mobility**, **Google Mobility**, **OpenSky**.

International Agreements:

- The biggest eight contributors (including EU) covered more than 70% of the worlds greenhouse gas emissions in 2017, and are therefore most relevant for the project.
- The scope was limited to their specific goals, as we had to create our own dataset from unformatted information.

To address our research question even better, we have already attempted to access weather data from NASA satellites, as we want to quantify the influence of precipitation on emission measurements. All datasets mentioned above are already accessible as a pandas dataframe in our repository. With these datasets, we feel confident to start implementing a model.

References

[1] Population figures for countries, regions (e.g. asia) and the world. `https://datahub.io/core/population#data`. Accessed: 2020-06-28.

[2] Bing covid-19 data. `https://www.bing.com/covid`. Accessed: 2020-06-28.

[3] Met Office. *Cartopy: a cartographic python library with a matplotlib interface*. Exeter, Devon, 2010 - 2015.

[4] Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university. `https://github.com/CSSEGISandData/COVID-19`. Accessed: 2020-06-28.

[5] The collaboration platform for api development. `https://www.postman.com/`. Accessed: 2020-06-28.

[6] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[7] Orange visual programming. `https://orange3.readthedocs.io/`. Accessed: 2020-06-28.

[8] NASA. *NASA APIs*. `https://api.nasa.gov/`, visited last: June 28, 2020.

[9] NOAA/ESRL Dr. Pieter Tans and Scripps Institution of Oceanography Dr. Ralph Keeling. *Trends in Atmospheric Carbon Dioxide*. `https://www.esrl.noaa.gov/gmd/ccgg/trends/data.html`, visited last: June 28, 2020.

[10] Oreggioni G. Guizzardi D. Muntean M. Schaaf E. Lo Vullo E. Solazzo E. Monforti-Ferrario F. Olivier J.G.J. Vignati E. Crippa, M. *CO2 and GHG Emission Reports*. `https://edgar.jrc.ec.europa.eu/overview.php?v=50_GHG`, visited last: June 28, 2020.

[11] IEA. *CO2 Emissions from Fuel Combustion*. `https://www.iea.org/subscribe-to-data-services/co2-emissions-statistics`, visited last: June 28, 2020.

[12] Madrid Government. *Air quality madrid 2001-2020*. `https://www.kaggle.com/miguelsxvi/air-quality-madrid-2001-2020?select=Interprete_ficheros_+calidad_+del_+aire_global.pdf`, visited last: June 28, 2020.

[13] India Government. *Air Quality Data in India (2015 - 2020)*. `https://www.kaggle.com/rohanrao/air-quality-data-in-india?select=city_day.csv`, visited last: June 28, 2020.

[14] Air Quality Historical Data Platform. *Air Quality Data*. `https://aqicn.org/data-platform/register/`, visited last: June 28, 2020.

[15] Apple. *COVID-19 Mobility Trends Reports*. `https://apple.com/covid19/mobility/`, visited last: June 28, 2020.

[16] Apple. *COVID-19 Community Mobility Reports*. `https://google.com/covid19/mobility/`, visited last: June 28, 2020.

[17] OpenSky Network. *OpenSky Network Flightlists*. `https://opensky-network.org/datasets/covid-19/`, visited last: June 28, 2020.

[18] Flightradar24. *Flightradar24 Air Traffic Statistics*. `https://flightradar24.com/data/statistics/`, visited last: June 28, 2020.

[19] European Commission. *Reducing emissinos from aviation*. `ec.europa.eu/clima/policies/transport/aviation_en`, visited last: June 28, 2020.