Relation between climate change and new diseases

Robin Jose Raju

June 2024

1 Question

In recent decades, we have been witnessing an alarming increase in emergence of new diseases in humans. Climate change could be a significant contributor for this trend. In this project, the main question I try to answer is to find out the relation between the climatic changes and new diseases found in human beings. I plan to study which all countries have high new disease outbreaks and compare it with the gases emitted by the country on the same year.

2 Data Sources

2.1 PRIMAP-crf

All countries are required to report their domestic emissions to United Nations Framework Convention on Climate Change (UNFCCC) in the Common reporting Format(CRF) on annual basis. This data is formatted to meet the IPCC 2006 guidelines and organised in a table containing all available countries and their corresponding greenhouse gas emissions.
This dataset gives a detailed view into the gases emitted by each available countries, along with amount, category from which the gas was emitted, and the year. This dataset contains emission data from 1986 to 2019.
The original data, which is freely available from UNFCCC website, is obtained and modified to meet the IPCC2006 guidelines. This helps in more standardized datasets and formats for helping future usage.

Structure & Quality
The PRIMAP-crf dataset is a tabular data in a .csv file which is formatted consistently with the PRIMAP2 interchange format.
We consider the data from 1996 to 2019 but some countries have failed to report the emissions of some gases for different years and this leads to not having the complete values. The dataset was last updated on 2021 with data recordings

till 2019, this makes the dataset 2+ years old and without the latest data of the last 4 years. For ease of usage, we use the total emissions of the country without considering from individual categories. We have also dropped the data from 1986 to 1995 as we don't have corresponding data with the diseases.

2.2 A global dataset of pandemic- and epidemic-prone disease outbreaks

This dataset contains new infectious diseases outbreaks collected from the Diseases Outbreak News(DONs). This dataset is product of a paper in which the researchers collected infectious disease outbreaks from DONs and Coronavirus Dashboard produced by World Heath Organization. The dataset contains information on diseases occurred over the period from 1996 to 2022 in 233 countries and territories around the world but we are using only the data till 2019 as the PRIMAP-crf contains data till 2019. The researchers have classified Africa, America and Asia as hot spots since they noticed high incidences of outbreaks in these regions.

The paper produces different layouts of the dataset from which 'Outbreaks.csv' is the one that we are interested in as they have all the new outbreaks and can be easily integrated to the dataset of domestic emissions.

Since the data is obtained directly from the WHO, modification on it is necessary to adapt the dataset for different use cases and the researchers have divided the datasets into different subsets of the main dataset. This helps us to choose the data that is relevant for the project and neglect other unwanted data.

Structure & Quality

From different subsets of the dataset, I choose the Outbreaks subset, which contains all required information like country code based on ISO 3166, year of the occurrence of the outbreak, name according to ICD-10 etc.

I have modified the dataset for a simpler and easier representation of the data, which will help in removing all unwanted columns that do not contribute significantly to the project and also truncated the data to contain only data from 1996 to 2019. All naming of diseases and country names follow the standard codes like ICD-10, ISO-3166 etc.

3 License

The 2 datasets that are used in this project are covered under an Open Data CC BY 4.0 license, which can be confirmed by going to the respective metadata links: PRIMAP-crf & A global dataset of pandemic- and epidemic-prone disease outbreaks.

The CC BY 4.0 license lets users to share and adapt the dataset to meet the requirements of the project. I will be citing the papers, datasets and the license on the final report and the results of the project will also come be released to the public for future usage.

4 Data Pipeline

The first attempt on the project was done using Jayvee but had to switch to Python, due to a limitation of Jayvee not supporting ".7z" file type in the ArchiveInterpreter. I have made use of the "py7zr", "requests" and "pandas" library for the data pipeline. In the project, I started of with getting the ".csv" file using "pandas" library for the first dataset. For the second dataset, I had to use "requests" and "py7zr" along with "pandas" library to download, extract and access the ".csv" file.

The next step would be to get the subset of the dataset that you actually need for the project. I started by selecting the required columns from the main dataset and creating a new data frame with those columns only rather than modifying the main dataset.

Since the PRIMAP-crf dataset have some missing values and its not easy to calculate a default value, I decided to drop those rows and also use the rows which gives the total emissions of the country rather than individual emissions from each industry of the country. Similarly, for the second dataset we truncated the rows so that we have only values till 2019. After all these operations, the resulting datasets are written into 2 separate ".sqlite" file using to_sql().

5 Result & Limitations

After successfull run of the data pipeline, the results are 2 ".sqlite" files named "PRIMAP-crf" and "diseases". The "diseases.sqlite" file contains 1458 rows of data with 'Country', 'iso3', 'Year', 'icd10n', 'icd104n', 'icd11l1', 'icd11l3' and 'Disease' as columns. For further information about the column names, please click here.

The "PRIMAP-crf.sqlite" contains 1367 rows of data with 'area(ISO3)', 'entity', 'unit', 'category (IPCC2006)', and then the years from 1996 to 2019 as columns. Both of the tables have datatype assigned to each column, according to their data, by pandas itself.

On primary analysis, the data seems to be in good shape and can be used for further study to see how the constituents of the atmosphere have affected the rise in new diseases. A weakness I noticed is that from a huge number of data only 1300+ of them turned out to be useful in both datasets, this is something that I would keep in mind for future.

Disclaimer : The main agenda of the project, to find out the relation between the gases emitted by a country and new diseases outbreaks reported in the country, may not be a direct implication for the final verdict. If the results of this project turns out to be a positive output, then more deeper studies in different levels should be required to support the verdict.