Name: Diana Jaganjac

Student Number: 13170192

Date: 24/02/2020

Academic Declaration: "I have read and understood the sections of plagiarism in the College Policy on assessment offences and confirm that the work is my own, with the work of others clearly acknowledged. I give my permission to submit my report to the plagiarism testing database that the College is using and test it using plagiarism detection software, search engines or meta-searching software."

Kaggle: Global Power Plant Database Analysis

Introduction:

The Kaggle Dataset I have decided to explore is the Global Power Plant Database which is provided by the World Resources Institute. I decided on this database as I am planning to complete my MSc thesis on greenhouse gas emissions detection by satellite imagery and believe that this dataset may provide useful for my future studies.

The dataset which I have chosen, is also accompanied with further data in an image (.tiff) format. Although I refer to this data which comprises satellite imagery, I focus on the Global Power Plant Database, which is contained within an excel file in (.csv) format. For the purposes of this report, I have decided to stick to this dataset, however a more extensive analysis would also require exploration of the (.tiff) files. Furthermore, the Kaggle competition requires participants to focus on Puerto Rico as the basis for analysis, however for my purposes, I will consider the entire Global Power Plant Database.

Brief Description:

The Global Power Plant Database forms up part of a selection of data which is provided for a Kaggle competition. The competition is an analytics prediction competition which sets out to explore alternatives for emissions factor calculations. Namely, are emissions factors calculable using remote sensing from satellite imagery. The other data provided from satellite imagery for this competition include: Global Forecast System imagery (NOAA, 2020), Global Land Data Imagery (NASA, 2020) and Sentinel Pollution Imagery (EU, 2020).

The purpose of the Kaggle competition, run by Data Science for Good (DS4G), is to research and analyse if there is a better way to model greenhouse gas emissions factors (see Graph. 1) from satellite imagery. Currently, emissions factors are measured using several different datasets, a few of which may include road transportation emissions, aviation emissions, public electricity usage, individual carbon footprint etc (Streets et al, 2013). This has proven to be useful in recent years for estimating greenhouse gas emissions, however, is also prone to uncertainty and inaccuracy of data, time-consuming data collection and overly general and unreliable quantitative modelling (DS4G, 2020). Another approach which is explored by this Kaggle competition, is to use satellite imagery to measure, calculate and predict greenhouse gas emissions factors with more accuracy, speed and reliability than traditional methods.

The Global Power Plant Database (GPPD) fits in to this selection of data as it enables the retrieval of the exact locations of power plants across the world. This is useful when analysed against pollution levels as it may help to pin-point and explain areas of increased

pollution. The GPPD is comprised of 24 dimensions which indicate the location, capacity, fuel type and the age of power plants globally. The database contains information for approximately 28,700 power plants geolocated to 164 countries across the globe and amasses data from 600+ sources of information (World Resources Institute, 2019). The GPPD provides data which becomes especially useful when combined with remote sensing techniques and applications. For instance, satellite imagery of local weather forecasts and pollution levels superimposed with the mapping of power plants based on their geo-referenced location provide unique and insightful inferences which could lead to emissions factor predictions (Diem and Comrie, 2002).

A rather pertinent limitation of satellite imagery is that it can be impacted by cloud cover. Cloud cover may conceal all or part of an image, therefore rendering it unusable for further analysis (Campbell, 1996). For this reason, DS4G recommend that competition participants use Puerto Rico as the basis for their analysis. As an island, Puerto Rico is less likely to be impacted from neighbouring atmospheric and weather systems. There is also the added benefit of it being easier to identify and isolate pollution attributable to power plants and energy generation (see Graph. 2).

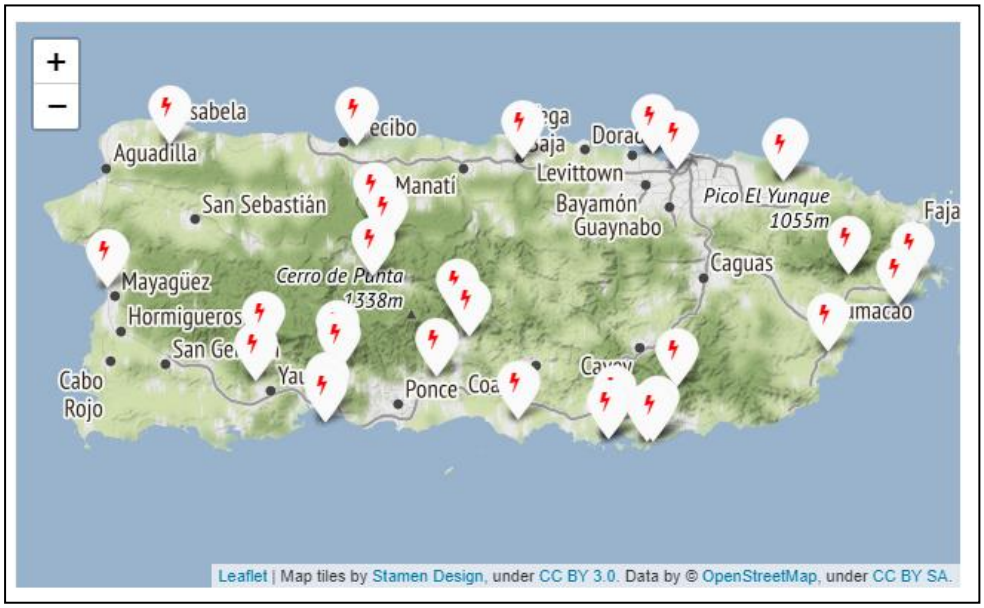
The general equation for emissions estimation is:

$E = A \times EF \times (1-ER/100)$

where:

- E = emissions;
- A = activity rate;
- EF = emission factor, and
- ER =overall emission reduction efficiency, %

Graph. 1: The equation for greenhouse gas emissions estimation (EPA, 2020).



Graph. 2: Map of Puerto Rico showing locations of all power plants. (Understanding the Data WIP, 2020)

Dimensional Analysis:

The Global Power Plant Database (GPPD) consists of 24 dimensions, and a total of 459, 441 out of a possible 717,840 data points. A good understanding of the following 6 dimensions, is necessary to understand how the data is distributed: capacity_mw, longitude and latitude, primary fuel, commissioning year and estimated_generation_gwh. These 6 dimensions can be further categorised as either qualitative or quantitative. For example, commissioning year and primary fuel give descriptive qualitative data, whereas longitude, latitude, capacity_mw and estimated_generation_gwh provide quantitative numerical data.

The qualitative data, commissioning year and primary fuel, provide essential information which assists the user to better understand the overall scope and relevance of the database. For example, the commissioning year dimension gives data in a year format (‘yyyy’) and ranges from 1896 to 2018. This indicates the timeframe of this dimension to be 120 years and provides the user with a good range of global power plants through time. However, there are data entries for this dimension which come as a decimal, for instance ‘1903.666667’ or ‘2002.154286’ (see Graph. 3). This is of course an issue, as we do not tend to depict years as decimals. In addition, this dimension provides 16,304 data-points out of a possible total of 29,910 which means that 13,606 data points are missing.

The primary fuel data dimension is qualitative and categorical and provides users with information on the primary fuel source used for electricity generation. This dimension consists of the following categories: Biomass, Coal, Cogeneration, Gas, Geothermal, Hydro, Nuclear, Oil, Other, Petcoke, Solar, Storage, Waste, Wave and Tidal, Wind. This data is useful as it provides the user with useful information which can help to attribute greenhouse gas atmospheric pollutants to certain areas where a power plant may be operating by burning coal, for instance. The data is complete and no data -points are missing, however the category 'other' may prove problematic.

The data dimensions, latitude and longitude, provide the geolocation of power points in decimal degrees and within the WGS84 coordinate system. Although technically these dimensions are provided as decimals, they serve a more descriptive purpose. For example, by using these data-points to plot the exact geo-referenced locations of power plants, the user can better understand the global scale of the database. Although I have not chosen it as one of my 6 data-dimensions, the country_long data dimension is also useful here as it provides information on which country a power plant belongs to. The two (or three) data dimensions are useful as they provide the user with information which is pertinent to the use of geo-referenced spatial imagery such as remote sensing satellite imagery. The latitude and longitude data dimensions are of good data quality, and there are no missing data-points.

The final two data dimensions, capacity_mw and estimated_generation_gwh provide essential quantitative data. The capacity_mw dimension gives the capacity of electrical generation in megawatts, and the estimated_generation_gwh gives the estimated annual electricity generation in gigawatts per hour for the year 2014. The capacity_mw dimension has no missing data-values, however there is an inconsistency in its data values. For instance, some are represented as decimals ('153.94') and other as whole numbers ('290'). There is also a lack of clarity regarding the timeframe of the dimension. For instance, is the capacity of electrical generation (mw) per second, minute, hour or year? This is particularly unclear when considering the range of the data values, which start at '1' and end at '22500'. Further clarification would be needed to improve the integrity and practicality of this data dimension.

The estimated_generation_gwh data dimension is comprised of 21,792 data-points, of a possible 29,910, meaning that there are 8118 missing values. The meta-data for this dimension also indicates that the dimension only provides estimates for the year 2014. This is particularly confounding, as the adjacent data dimensions, 'generation_gwh_2013' to 'generation_gwh_2017' are too partially complete. There are instances where these data dimensions are empty, and yet the estimated_generation_gwh contains a value (see Graph. 4) Hence, the integrity and quality of the estimated_generation_gwh is highly questionable and is difficult to trust as being accurate and reliable.

In conclusion, it is evident that the Global Power Point Database (GPPD) does contain data dimensions which are useful and reliable. For instance, the primary fuel, longitude, latitude and capacity_mw data dimensions do not contain any missing values and are generally sound. Despite this, the primary fuel and capacity_mw data dimensions are of questionable integrity and may cause some confusion in understanding how data are distributed. On the other hand, the commissioning year and estimated_generation_gwh data dimensions are incomplete and hence, to some extent, unreliable. Nevertheless, I would argue that that the GPPD is generally useful and that the database can contribute positively to larger studies particularly those where satellite remote sensing techniques are used as a tool for analysis.

	A	B	C	L	M	N
1	country	country_long	name	commissioning_year	owner	source
39	USA	United States of America	Madison Dam	1906.5	NorthWestern Energy	U.S. Energy Information Administration
40	USA	United States of America	Bishop Creek 4	1906.756098	Southern California Edison Co	U.S. Energy Information Administration
41	USA	United States of America	Georgetown Hydro	1907	Public Service Co of Colorado	U.S. Energy Information Administration
42	USA	United States of America	White River (WI)	1907	Northern States Power Co - Minnesota	U.S. Energy Information Administration
43	USA	United States of America	Fall Creek	1907.761905	PacifiCorp	U.S. Energy Information Administration
44	USA	United States of America	Whiting	1907.941176	Consolidated Water Power Co	U.S. Energy Information Administration
45	USA	United States of America	Wainiha Hydro	1908	McBryde Sugar Co Ltd	U.S. Energy Information Administration
46	USA	United States of America	Deer Creek PH	1908	Pacific Gas & Electric Co.	U.S. Energy Information Administration
47	USA	United States of America	Cottonwood	1908.5	Los Angeles Department of Water & Power	U.S. Energy Information Administration
48	USA	United States of America	Bishop Creek 2	1908.945205	Southern California Edison Co	U.S. Energy Information Administration
49	USA	United States of America	Lower Tule River	1909	Southern California Edison Co	U.S. Energy Information Administration
50	USA	United States of America	Johnsonville Dam	1909	Erie Boulevard Hydropower LP	U.S. Energy Information Administration
51	CHE	Switzerland	Ackersand 1	1909		Swiss Federal Office of Energy
52	USA	United States of America	Croton	1909.101124	Consumers Energy Co	U.S. Energy Information Administration
53	USA	United States of America	Little Quinnesec Falls	1909.488889	Northbrook Wisconsin LLC	U.S. Energy Information Administration
54	USA	United States of America	Turners Falls	1909.807018	FirstLight Power Resources Inc. - MA	U.S. Energy Information Administration
55	USA	United States of America	Biron	1909.928571	Consolidated Water Power Co	U.S. Energy Information Administration
56	USA	United States of America	Domtar Paper Compa	1910	Domtar Paper Company Rothschild	U.S. Energy Information Administration
57	USA	United States of America	Snake Creek	1910	Heber Light & Power Company	U.S. Energy Information Administration
58	USA	United States of America	Pioneer	1910	PacifiCorp	U.S. Energy Information Administration

Graph. 3: Global Power Plant Database showing variation in recording of commissioning_year dimension (column L), (World Resources Institute, 2019).

	S	T	U	V	W	X
1	generation_gwh_2013	generation_gwh_2014	generation_gwh_2015	generation_gwh_2016	generation_gwh_2017	estimated_generation_gwh
2						
3						
4						89.13207547
5						1650.59399
6						1980.712788
7						16.5059399
8						79.22851153
9						82.52969951
10						825.2969951
11						0
12						2152.249819
13						293.8648791
14						2317.807497
15						413.8941959
16						1862.523882
17						1862.523882
18						1208.571052
19						4966.730351
20						1730.077739
21						298.0038211
22						2483.365176

Graph. 4: Global Power Plant Database showing empty data dimensions (columns S – W) yet estimated_generation_gwh is non-empty (column X), (World Resources Institute, 2019).

References:

Campbell, J., 1996. Introduction to Remote Sensing. London: Taylor & Francis.

Diem, J., Comrie, A., 2002. Predictive mapping of air pollution involving sparse spatial observations. Environmental Pollution, 119(1), 99 – 117.

DS4G, DS4G: Environmental Insights Explorer, available online at https://www.kaggle.com/c/ds4g-environmental-insights-explorer. Last accessed 25/02/2020.

EPA, Air Emissions Factors and Quantification, available online at https://www.epa.gov/air-emissions-factors-and-quantification/basic-information-air-emissions-factors-and-quantification#About%20Emissions%20Factors. Last accessed 25/02/2020.

EU, Sentinel-5P NRTI NO2: Near Real-Time Nitrogen Dioxide, available online at https://gee.stac.cloud/22Xh3VjjQHaEAfFGXRb4ko1DnpuvqpFCT2qByfnr7wb?t=bands. Last accessed 25/02/2020.

NASA, Global Land Data Assimilation System, available online at https://ldas.gsfc.nasa.gov/gldas/. Last accessed 25/02/2020.

NOAA, Global Forecast System, available online at https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php. Last accessed 25/02/2020.

Streets, D., et al., 2012. Emissions estimation from satellite retrievals: A review of current capability. Atmospheric Environment. 77, 1101, 1142.

Understanding the Data WIP, Understanding the Data WIP, available online at https://www.kaggle.com/parulpandey/understanding-the-data-wip. Last accessed 25/02/2020.

World Resources Institute, Global Power Plant Database, available online at http://datasets.wri.org/dataset/globalpowerplantdatabase. Last accessed 25/02/2020.