CDP Full GHG Emissions Dataset

2019 Summary

1 Introduction

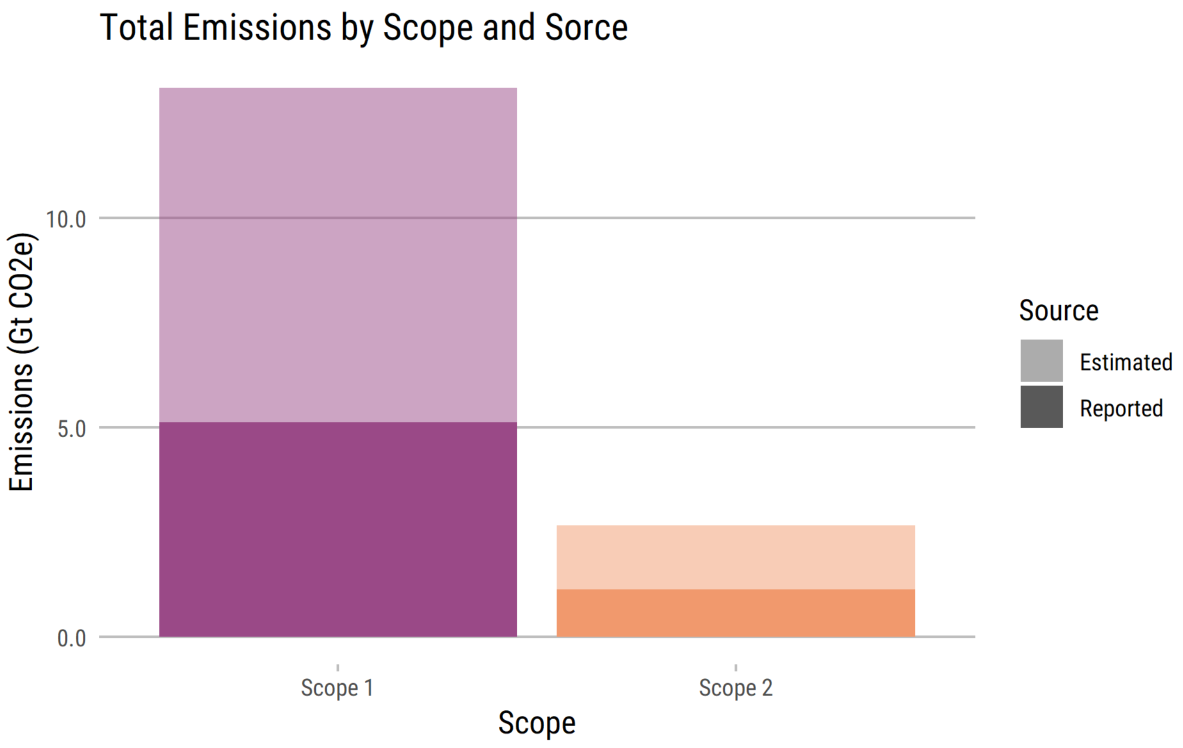
There is a growing appetite from investors for high quality and complete corporate greenhouse gas (GHG) emissions data. Since 2015, CDP have compiled an annual dataset covering companies in the highest emitting industries.

CDP owns and controls one of the largest databases of primary corporate environmental data in the world. This database is leveraged to produce energy consumption and emissions estimates for companies that do not disclose. It also provides the means to improve the quality of company disclosures by identifying anomalies and engaging with the companies to resolve them.

The dataset includes data or estimates over 5000 companies, whose Scope 1 emissions alone equated to over 35% of global greenhouse gas emissions in 2018. The sample consists of companies in the MSCI All-Country World Index (ACWI), as well as the highest emitting companies not included in this index. For each company, the dataset contains reported or estimated values for the following datapoints:

- Absolute Scope 1 and Scope 2 GHG emissions
- Total fuel consumption and purchased steam, heat, electricity & cooling (SHEC) use
- Emissions revenue intensity (Scope 1 + 2 per million units of revenue in US dollars)
- For companies in high emitting sectors, emission intensities are also provided per unit of physical output.

This document is an overview of data cleaning and modelling methods used to enhance self-reported data from companies. For more detailed technical descriptions of these processes, please refer to the technical annexes available on CDP's website.



2 Methods

2.1 Cleaning

CDP cleans reported emissions data for two reasons: (1) to improve the accuracy and reliability of the data, and (2) to improve the robustness of the statistical model estimates, which are derived from the dataset.

Internal consistency checks are used for every CDP responding company to verify that each data point aligns with the other information that the company has reported. External consistency checks rely on information from other data sources to try to reconcile the companies' disclosed emissions with what is reported elsewhere.

First, the checks for internal consistency are used to remove data points that may have been misreported. Large outliers can distort statistical models and consequently these are removed. Once this has been done, the models can be used to identify companies which are potentially misreporting data. Any data points that the models identify as outliers are then investigated in detail by thoroughly reviewing key data points across the companies' questionnaire responses.

If a datapoint appears to have been misreported, and the company is one of the top 200 highest emitters from the previous year, they are contacted for clarification. For missing or possibly misreported data, an estimate is provided alongside the reported value.

2.2 Modelling

2.2.1 Bottom-up Models

Bottom-up modelled estimation is based on the combining of physical activity indicators (tonnes, barrels, kilometers, etc.) and their associated emission factors. Because such indicators relate directly to the emitting activity, the bottom-up estimates are expected to have a narrower margin of error than the statistical estimates. Activities and products are accounted for individually, so the method is limited to homogenous sectors. Homogenous sectors are so-called because they are structured around one, or a low number of, processes and products. Such structures are most common in the upstream extraction, production and conversion industries, which is why the method accounts for the most energy- and emissions-intensive sectors.

Six homogenous sectors are bottom-up modelled for corporate energy and emissions estimation. Asset-level data is utilized for six of the seven sectors.

Sector activity	Data points modelled
Coal mining	S1, S2, S3 (Use of Sold Products), Fuel, SHEC*
Oil and gas extraction	S1, S2, S3 (Use of Sold Products), Fuel, SHEC*
Petroleum refining	S1, S2, S3 (Use of Sold Products), Fuel, SHEC*
Electric power generation	S1, S2, Fuel, SHEC*
Steel production	S1, S2, Fuel, SHEC*
Cement production	S1, S2, Fuel, SHEC*

*SHEC – Purchased energy in the form of steam, heat, electricity and cooling

2.2.2 Intra-company Models

Where a company has provided some but not all of the data points covered under this project, CDP has used the available data to calculate the remaining emissions and energy figures. This is done in the following cases:

- *IEA emissions factors applied to reported data:* if the company has provided a figure for SHEC in MWh but not location-based Scope 2 emissions, then the IEA national level grid emissions factor has been applied to SHEC to calculate the emissions figure.
- *Standard factors applied to reported data*: if the company has provided a figure for fuel use in MWh but not Scope 1 emissions, then a standard emission factor has been applied to the energy consumption in order to calculate the Scope 1 emissions (or vice-versa). The emission factor used is based on the fuels commonly used by other companies in the sector (according to data reported to CDP).
- *Split emissions found in company filings:* where companies have reported a total GHG figure in their filings, CDP has split this total figure into Scope 1 and Scope 2 emissions using the statistical models to estimate the proportion of Scope 1 and Scope 2 emissions for that company.
- *Previous year's data carried forward:* where companies have reported data for previous year, but not the current year, the previous value is carried forward. This assumes that the variability in reported data between companies is larger than the year-on-year variability of a single company.

2.2.3 Statistical Models

Multi-variate regression models are used to estimate data points in sectors where no bottom-up estimates are produced, with company revenue and activity classification as the predictor variables. The revenue data is extracted from Bloomberg and is broken down by business activity according to the Bloomberg Industry Classification System (BICS), allowing estimates for companies with more than one activity. Business activities in the models are classified under the CDP Activity Classification System which has been optimized to provide robust estimates for this project.

These regression models are used for calculating Scope 1 emissions, SHEC and fuel. For calculating the location-based Scope 2 emissions, the SHEC value is multiplied by the national grid emissions factor of the company's country of operations. Regional revenue breakdown data is taken from Bloomberg so that international companies are accounted for. Several generalizing assumptions are made so that the models can be applied to all regions and activities. For this reason, the bottom-up estimates have been used wherever possible and the statistical results leveraged to fill remaining gaps.

3 Data Sources

3.1 Prioritization

Sources of data are prioritized by order of their perceived accuracy, as well as the transparency of the companies' calculation methodologies. CDP uses reported data wherever possible, only using models to fill gaps in the data and to provide estimates as a comparison to any misreported data. In order of reliability, the dataset sources are:

- **Data Check data:** For the top 200 highest emitters in the previous year, CDP contacted the companies to clarify potential issues with the reported data. This is considered the most reliable source of data because CDP analysts are in direct contact with the company and will check the data as it is submitted.

- **CDP data:** The CDP questionnaire requires companies to provide information about their calculation methodology, such as accounting boundary, omissions, energy use, breakdowns, etc. All of this information is used to validate the emissions reported, and so CDP data is chosen over company filings data wherever possible. This is not to say that CDP data is more reliable than data reported in company filings when it is reported, but that as all CDP data goes through a much more rigorous cleaning process than the CSR data, it is treated as more reliable.

- **Company filings data:** Data in company filings is harder to clean because companies often neglect to mention the reporting boundary and methodology used in their calculations. If information is provided, then it may be unclear or incomplete. This means that it is often harder to understand why a reported value differs from our expectations. Data from this source is used only occasionally.

- **IEA emissions factors applied to reported data:** This is done when a company has provided SHEC data, but no location-based scope 2 data. The reliability of the calculation is limited by the uncertainty and level of granularity in the emissions factors used.

- **Split emissions found in company filings:** these emissions figures are based on reported data but still rely on the statistical modelling to divide the total figures between Scope 1 and Scope 2 emissions. This means that they are less accurate than other sources of reported data, but still more accurate than the estimates alone.

- **Bottom-up model estimates:** these sector specific models use methodologies that are often similar to those used by the companies themselves to calculate their emissions using data disclosed by that company in 2016, and so they are much more robust than statistical models.

- **Regression estimates (gamma GLM):** The gamma generalized linear model (GLM) is the name of the family of regression models used for these estimates. The estimates from these models are calculated using self-reported GHG data, which requires cleaning, and with assumptions required to accommodate data constraints.

- **Standard emissions factors applied to estimated data:** This is used to estimate the location-based Scope 2 emissions figures. It takes the IEA national level grid emissions factors and multiplies it by the estimated SHEC value to provide an estimate for the emissions figure. These estimates are assumed to be the least reliable because there is uncertainty coming from the emissions factor used, and from the regression models.

3.2 Data Reliability Score

Using the information above, numerical scores from 1 to 7 have been applied to the data to give an indication of reliability, where 7 is the highest score. The quality of the regression models varies for different activities and depends on the consistency of the data reported by other companies in that sector. Consequently, different sectors have different quality flags for the regression models. Only self-reported data submitted to CDP that has passed the quality review receives a 7.