CDP Full GHG Emissions Dataset

2024 Summary

1 Introduction

There is a growing appetite from capital markets 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.

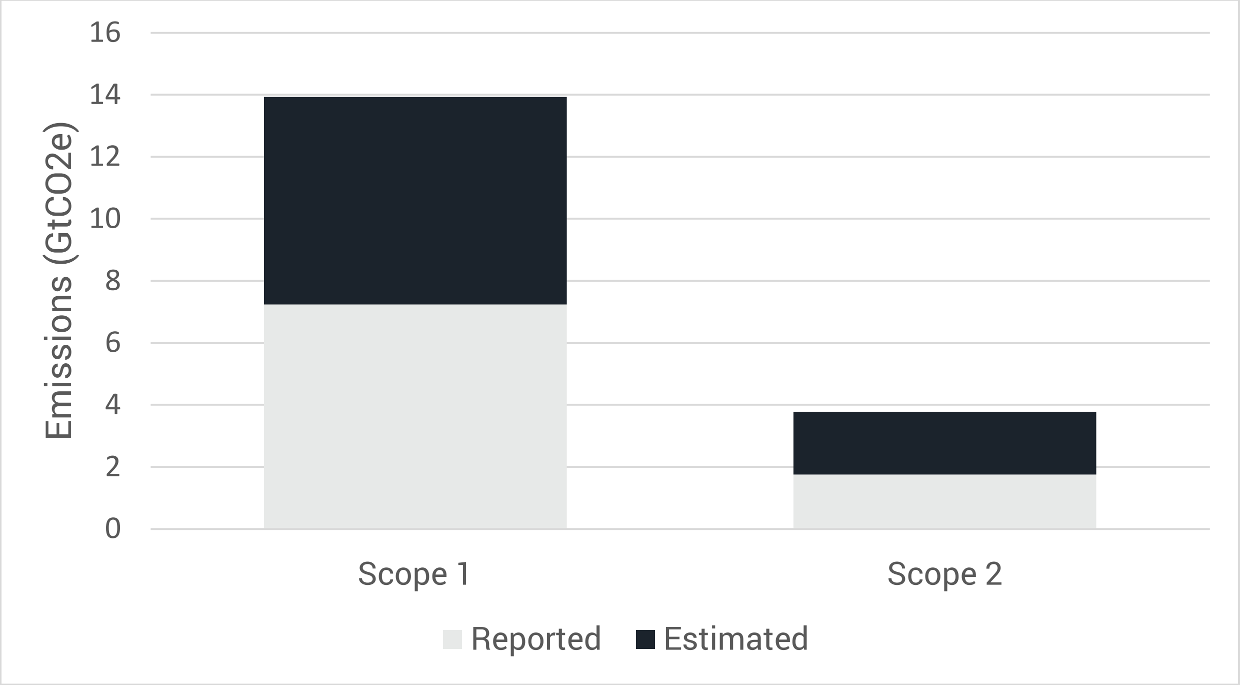
CDP owns one of the largest databases of primary corporate environmental data in the world. This database is leveraged to produce energy consumption and emissions estimates for requested companies that do not disclose through CDP. It also provides the means to improve the quality of company disclosures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

The dataset includes reported and estimated figures for approximately 14.6k companies. For each company, the dataset contains reported or estimated values for the following datapoints:

- Absolute Scope 1, Scope 2 and Scope 3 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)

This document gives an overview of the data cleaning and modelling methods used by CDP 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.

Total emissions by scope and source



2 Methods

2.1 Consistency & accuracy checks

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

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 exclude data points that may have been misreported. Large outliers can distort statistical models and consequently these are removed. Next, the models can be used to identify companies which are potentially misreporting data. In the final output no reported data is removed. Instead, for missing or possibly misreported data, an estimate is provided alongside the reported value.

2.2 Modelling

2.2.1 Intra-company models

Where a company has provided some, but not all data points covered in this dataset, 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:** for scope 2, 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. We would use the reported regional SHEC breakdown if available, otherwise, we use regional revenue as a proxy to calculate regional SHEC.
- **Incomplete reported data:** for SHEC, if a company reports non-renewable and renewable energy consumption but doesn't report the total, we sum the two values together to calculate the total SHEC.

2.2.2 Statistical models

Multi-variate regression models are used to estimate data points, using company revenue and activity classification as the predictor variables. The revenue data is broken down by business activity according to CDP Activity Classification System data, 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.

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:

- **CDP data:** The CDP questionnaire requires companies to provide information including activity data, revenue, energy use, calculation methodology, breakdowns, etc. All this information is used to validate the emissions reported, and so CDP data is chosen over company filings data wherever possible. Although the raw data disclosed to CDP is not necessarily more reliable than data reported in company filings, CDP data goes through a more rigorous cleaning process than the data included in companies' Corporate Social Responsibility (CSR) reports and is therefore deemed more reliable.
- **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.
- **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 Quality Score

All reported and estimated emissions figures are accompanied by a data quality score from the Partnership for Carbon Accounting Financials (PCAF). Scores from 1 to 4 are applied (PCAF score 5 is not applicable to our data) to the data to give an indication of reliability, where 1 is the highest score for verified reported emissions. Our physical activity model would receive a PCAF score 3 whilst our regression models would receive a PCAF score 4.

Below are the PCAF scores and how they map to data quality within the Full GHG Emissions Dataset:

PCAF Score	CDP Description
1	**Verified emissions** of the company are available
2	**Unverified emissions** calculated by the company are available. Emissions are calculated using primary physical activity data of the **company's energy consumption** and emission factors specific to that primary data. Relevant process emissions are added.
3	Emissions are calculated using primary physical activity data of the **company's production** and emission factors specific to that primary data.
4	Emission factors for the sector per unit of revenue are known (e.g., tCO2 e per euro or dollar of revenue earned in a sector).
5	N/A

4 Key Differences from 2023 Dataset

The key differences between the 2023 and 2024 datasets are as follows:

- Increased sample size from 13.5k to 14.6k (~8% increase).
- Inclusion of small and medium-sized enterprises (SMEs).
- Company revenue and activity data sourced directly from companies through questionnaire set-up (revenue data has also been sourced from S&P, as was the case in 2023).
- Remaining physical activity estimates generated from asset- or company-level production data have been phased out.
- Automation of data cleaning to ensure scalability (manual flagging of data points still occurs on occasion).