CDP
DISCLOSURE INSIGHT ACTION

CDP full GHG emissions dataset

2023 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.

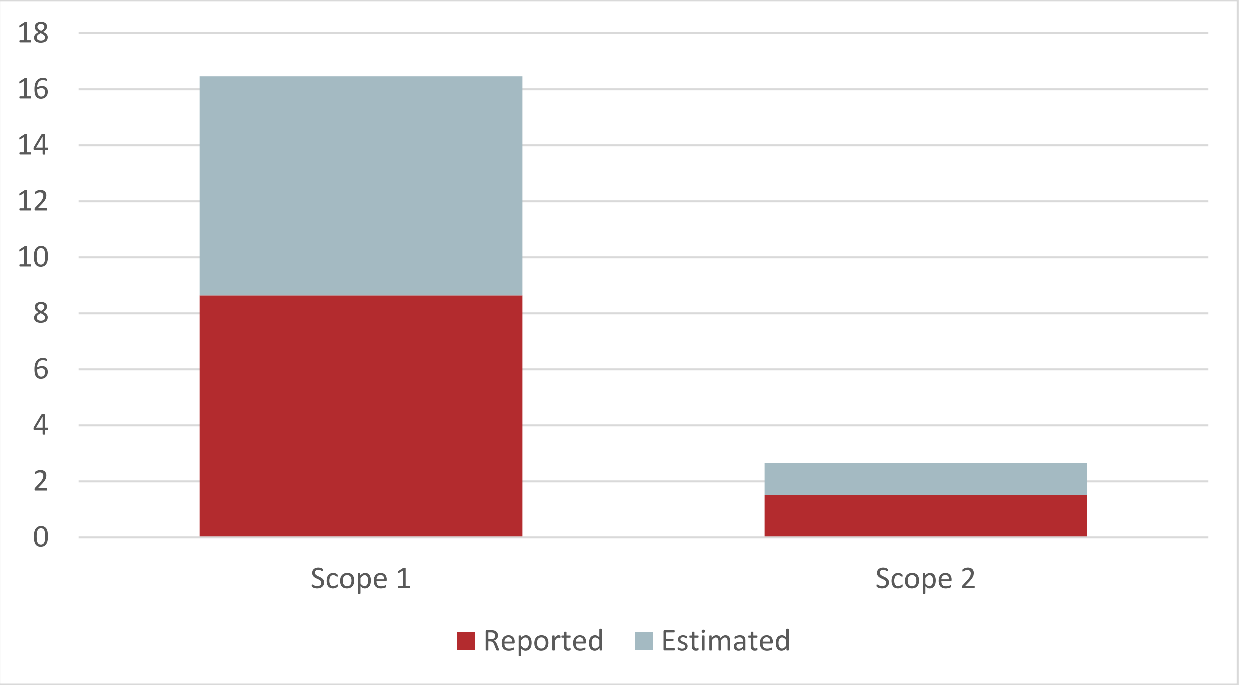
CDP owns and controls one of the largest databases of primary corporate environmental data in the world. This database is leveraged to produce energy consumption and emissions estimates for companies that do not disclose. It also provides the means to improve the quality of company disclosures by identifying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

The dataset includes reported and estimated figures for approximately 13.5k companies, whose Scope 1 emissions alone equated to over 28% of global greenhouse gas emissions in 2022. 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)

This document is 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 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.

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. Next, 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. For missing or possibly misreported data, an estimate is provided alongside the reported value.

2.2 Modelling

2.2.1 Physical activity models

Physical activity modelled estimation is based on the combining of physical activity indicators (tonnes, barrels, kilometres, etc.) and their associated emission factors. Because such indicators relate directly to the emitting activity, the physical activity 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 as upstream extraction, production and conversion industries. In the Full GHG Emissions Dataset we focus on four of the most energy intensive sectors.

Four homogenous sectors are modelled for corporate energy and emissions estimation.

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, S3 (Fuel- and Energy Related Activities), 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. We would use the regional breakdown if available otherwise we use regional revenue as a proxy to calculate regional SHEC.
- **Incomplete reported data:** If a company reports non-renewable and renewable energy consumption but doesn't report the total, we would automatically calculate the total SHEC.

2.2.3 Statistical models

Multi-variate regression models are used to estimate data points in sectors where no physical activity estimates are produced, with company revenue and activity classification as the predictor variables. The revenue data is extracted from S&P 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 S&P so that multiple international operations are accounted for. Several generalizing assumptions are made so that the models can be applied to all regions and activities. For this reason, the physical activity 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:

- **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. Although the raw data disclosed to CDP is not necessarily more reliable than data reported in company filings, CDP data goes through a much more rigorous cleaning process than the data included in companies' Corporate Social Responsibility (CSR) reports and is therefore seen 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.

- **Physical activity model estimates:** these sector specific models that use a company's physical activity to estimate their emissions. Because these estimates are based on a company's production of a certain product, 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 Quality Score

We use a data quality score from the Partnership for Carbon Accounting Financials (PCAF). Scores from 1 to 4 have been applied (we don't use PCAF score 5) 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:

PCAF Score	PCAF Description
1	Outstanding amount in the company and EVIC are known. **Verified emissions** of the company are available
2	Outstanding amount in the company and EVIC are known. **Unverified emissions** calculated by the company are available.
	Outstanding amount in the company and EVIC are known. Reported company emissions are not known. 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	Outstanding amount in the company and EVIC are known. Reported company emissions are not known. Emissions are calculated using primary physical activity data of the **company's production** and emission factors specific to that primary data.
4	Outstanding amount in the company, EVIC, and the **company's revenue** are known. 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	Outstanding amount in the company is known. Emission factors for the sector per unit of asset (e.g., tCO2 e per euro or dollar of asset in a sector) are known.
	Outstanding amount in the company is known. Emission factors for the sector per unit of revenue (e.g., tCO2 e per euro or dollar of revenue earned in a sector) and **asset turnover ratios** for the sector are known.