CDP full GHG emissions dataset

Technical Annex I: Data Cleaning Approach

Contents

1 Introduction

The Full GHG Emissions Dataset provides CDP's investor members and other stakeholders with the most up-todate, accurate and comparable corporate GHG emissions and energy-use data. This is one of a series of documents outlining how the raw reported data is enhanced. All are available on CDP's website.

- CDP Full GHG Emissions Dataset: Summary 2023
- Technical Annex I: Data Cleaning Approach
- Technical Annex II: Physical Activity Modelling
- Technical Annex III: Statistical Framework
- Technical Annex IV: Scope 3 Overview and Modelling

This document provides an overview of the methods used to clean and validate the data reported by companies. These methods leverage a statistical framework developed by CDP's Data Analytics team. For a detailed description of this framework, refer to *Technical Annex III: Statistical Framework*.

1.1 Data Cleaning

Data cleaning is the process of identifying invalid or inaccurate records within a dataset. For reported emissions data, this is useful because it yields a more reliable dataset for analysis, and because it permits the development of more reliable statistical models, which are used to fill gaps. At a high level, the accuracy and reliability of reported data is evaluated in the following ways:

- *Internal consistency:* Comparing a reported value against the other data reported by the company in question. This includes looking at historical data and other data points from the company's CDP response to assess the reliability of the data point.

- *External consistency:* Checking that the data point is in line with the values reported by other similar companies. These checks are carried out using several statistical and physical models based on data collected from annual reports. In some cases, reported data is also compared against other external data sources including company filings.

The cleaning approach described here focuses on emissions from Scopes 1 & 2 of the Greenhouse Gas Protocol. There are four key data points involved; *Scope 1 emissions* are typically the result of *fuel combustion*, whereas *Scope 2 emissions* are the result of purchased the result of *purchased energy* from a third party. The CDP Climate Change questionnaire[1] accounts for four types of purchased energy: steam, heat, electricity and cooling (SHEC).

1.2 Modelling

In this project, modelling serves two purposes: it allows anomalous data points to be identified and cleaned, as described in the last section, and it provides estimates for missing or less-reliable data points. The models used for cleaning the data reported to CDP are multi-variable regression models, using a company's activity classification and revenue to estimate their emissions. For

[1] The Climate Change questionnaire and guidance is available here.

certain key sectors, more precise physical activity estimates are produced using raw activity data – refer to *Technical Annex II Physical Activity Estimation Methodology* to learn more. The statistical models are only as reliable as data that is used to build them. To reach a workable solution within the constraints on data availability, several assumptions are made. These are discussed at length in *Technical Annex III: Statistical Framework*, a summary is provided here:

- For every CDP activity the revenue is directly proportional to emissions
- All companies engaged in each CDP activity produce reasonably similar products
- All companies engaged in each CDP activity sell these products at a reasonably similar price
- All companies engaged in each CDP activity produce these products in a reasonably similar way
- All companies in each CDP activity report their emissions using similar accounting practises
- The variation across regions is smaller than the variation across different sectors (apart from Scope 2 location-based emissions).

The cleaning and modelling processes complement each other; unreliable data points are removed from the model training data, and observations that deviate disproportionately from their modelled estimate are investigated to check their reliability.

2 Common Sources of Error

2.1 Boundary and Consolidation Issues

Emissions boundary change

Companies can consolidate their emissions using different reporting boundaries, the three most common are Financial Control, Operational and Equity Share. In many cases a company's reported emissions will be very different across each of these three boundary choices. When companies have switched their boundary from one year to the next, the emissions have been known to change by as much as 50%.

Revenue boundary ≠ emissions boundary

Another common issue is the case where a company revenue boundary is not the same as their emissions boundary. For example, companies can choose to report at subsidiary level, rather than at group level. The consolidated revenue from the subsidiaries will be observed by the model, and so the group company's reported emissions will be much lower than the 'expected' value.

Similarly, a company can earn revenue from parts of its operations that don't fall within its emissions reporting boundary. Where a company is receiving revenue from a minority stake in an asset, they may still report their emissions using an Operational/Financial Control boundary. This would result in a significantly lower revenue intensity for that company when compared to a peer reporting with an Equity Share boundary.

Rental companies

Companies that lease/rent assets to a third party should account for the emissions from those assets during their use under Scope 3 up/downstream leased assets, some companies report these emissions under Scope 1, while others don't at all. This problem is most pronounced in property companies which are not allowed or able to collect, electricity and/or fuel use data from their tenants, and therefore only report the emissions from their offices

Entity mismatch

Corporate ownership structures are complicated, and the legal entity covered in the CDP response may not match the listed legal entity covered in the income statements. There are countless examples of publicly owned group holding companies with wholly owned private operating companies (and vice versa). The listed group company's CDP response may not include the operating companies, but the group company's financial reports may show revenue from the operating companies. This would be a case of entity mismatch where the CDP response is provided under a different reporting boundary to the income statements.

2.2 Omissions/Exclusions

The CDP questionnaire allows companies to exclude parts of their business provided they disclose an explanation for the omission, and an indication of its significance. Many new disclosers are unable to collect all the necessary data for a complete disclosure in their first year and so they are forced to exclude parts of their business from their disclosure. Some companies disclose these exclusions, but many do not. In many cases where the model estimates are higher than the reported value it is because the company is only reporting for a small part of their business.

Where an omission or exclusion has been identified the analyst must decide whether the excluded parts of the business could make up a material portion of the total emissions for that company. This issue is very subjective, and it is rarely possible to accurately estimate the size of the omitted emissions; as a rule of thumb 5% of the total emissions is deemed to be significant.

If the omission is deemed to be material, then the data point is excluded from the model training dataset and is flagged in the final output.

2.3 Data Entry Error

Human error cannot be discounted when dealing with manually entered data. Users may present the data in the wrong format or in wrong units. For example, a common human error is reporting data with incorrect units, such as joules as opposed to megawatt hours, or simply kilo tonnes as opposed to tonnes etc. These cases are often most easily characterised by the large residuals and implied emissions factors that are 1000 times too big/small. CSR reports can sometimes be used to verify this sort of error because the company is more likely to specify the units, they use in its CSR report. In some cases, companies will copy the values directly across from their CSR report without converting the units.

2.4 Auto-Generation

As electricity generation becomes less centralised, more and more companies are generating their own electricity. This means that companies that have decided to generate their own

electricity will have lower Scope 2 emissions and purchased energy figures than their peers. If this generation comes from a non-renewable source like a diesel generator, then the company will have unusually high Scope 1 emissions and fuel consumption figures. The models do not explicitly account for the amount of auto-generation a company uses. Consequently, this information can be used to help clean the data and explain to end users of the data why a company's emissions figures seem different to their peers.

2.5 Classification Errors

Fundamentally the models are being used to compare each company with the data reported by other companies in the same sector, adjusting for each company's size and location. If the training data contains companies that have been categorised incorrectly then the model will try and compare it with the wrong group of companies. CDP uses its own classification system[2] to classify companies' activities; the raw data used to classify companies comes from the revenue breakdown data in their annual reports, and this data is sourced from Bloomberg. The granularity of the revenue breakdown data is variable and, at times, insufficient. The companies in this sample are highly diversified, with many activities, but some companies only provide a brief summary of these activities in their company filings. This results in classification error where the revenue for a company is not being correctly allocated.

3 Flagging data reliability issues

Once a data reliability issue has been identified, the data point is flagged so as to not mislead the user of the dataset. The flagged data points are also removed from the model sample, so that these values do not influence the estimates produced.

There are also cases where it is appropriate to remove a data point from the model sample without flagging it as a reliability issue. For instance, if the company violates some of the modelling assumptions without misreporting the data, then the company's emissions figures may distort the model. This does not mean that there is anything wrong with the company's disclosure, so the reported figures are displayed without a flag.

The table below summarises the common sources of error and the actions that are taken accordingly:

Source of error	Action Taken
Boundary & consolidation issues: A company is only disclosing emissions for a part of its business	All data points are flagged and removed from the model sample
Omissions and exclusions: The company's inventory is incomplete (e.g. data isn't available for a certain region/department)	Data points are re-scaled if possible, otherwise they are flagged and removed from the model sample

[2] The CDP Activity Classification System has been developed with this project in mind. It is made up of 208 Activities, which are grouped into 60 Activity Groups, and further grouped into 13 Industries (see appendix).

Data entry error: The reported data is in different units, e.g. reporting kilo tonnes as opposed to tonnes	An appropriate scaling factor is applied
Classification error: The company has been misclassified under the CDP Activity Classification System	The classification is corrected accordingly
The reported emissions data is inconsistent with the reported energy use data with no clear explanation	The data point is flagged and removed from the model sample
Where a data point or company is having an unduly large skewing effect on the model	The data point is removed from the model sample, and flagged if appropriate
Where a data point falls outside of the 99% confidence interval for the models	The data point is investigated, removed from the model training data and flagged, if appropriate.
Where a company uses an incorrect methodology for a specific Scope 3 category, as outlined in The GHG Protocol Corporate Accounting and Reporting Standard.	The methodology is investigated and the company's explanation is checked. If there is no reasonable justification for using an incorrect methodology, the data point is flagged and removed from the model sample.

4 Checks and Tests

4.1 Checks for Internal Consistency

CDP uses a range of different checks for internal consistency, outlined below:

- *Historical medians:* these can be used to spot sudden step changes in reported emissions figures. If a data point is far larger or smaller than its historical median, then this is likely to be the result of either a change in calculation methodology, or the result of some large corporate change, e.g. mergers and acquisitions activity.

- *Implied emissions factors:* the implied emissions factor used by the company can be calculated by dividing the reported Scope 1 or Scope 2 emissions figure by the fuel or SHEC value respectively:

    - Most fuels release between 0.1 and 0.5 tonnes of CO2e per MWh of energy
    - The IEA World Energy Statistics average electricity grid emissions factor is around 0.6 tonnes per MWh and ranges between around 0.1 in Norway to 1.1 in South Africa.

  Any company with an implied emissions factor outside of these ranges could be assumed to have made some sort of error with either their emissions figure and/or their energy figure in MWh.

- *Company responses to other questions:* data from several questions in the CDP questionnaire can be used to clarify possible data issues. This includes the identification of boundary issues and exclusions described previously.

- *Gross emissions vs. sum of breakdowns:* companies report both total emissions values and breakdowns (e.g. by country, type of fuel etc). If these data points are not consistent with each other, it may indicate that the data has been misreported.

4.2 Checks for External Consistency

There are two primary external consistency checks, outlined below:

- *Regression models:* the regression models provide a framework for comparing the emissions of companies of all different sizes and sectors, using revenue breakdown data and activity classification data.

- *CSR and annual reports, and company websites:* comparing data across different reported sources can be useful to understand the credibility of the reported data.

5 Cleaning Approach

It is important to make sure that the statistical models are reasonably robust before trying to use them to identify outliers and misreported data. Consequently, CDP's data analysts clean the data in order to create robust statistical models. The four stages of the cleaning process are outlined below:

5.1 Stage 1: Checks for Internal Consistency (Applied to all Data)

Using the checks for internal consistency described in the previous section, removing all data points that fail these tests before running the model.

5.2 Stage 2: Remove Data Points with Large Leverage (Applied to all Data)

To ensure that the models are as stable as possible during the cleaning process it is important to remove the data points which are having a disproportionate skewing effect on the models. These can be identified as the data points with large *Cook's Distances*.

This stage is where the improvements to the models are made. The basic process is:

1. Run models
2. Rank the data points according to their *Cook's Distances*
3. Remove data points with large *Cook's Distances* from the model training data
4. Investigate each one to establish whether the data point has been misreported, if so then it also needs to be flagged as a data quality issue for investors
5. Establish if other data points reported by the company are affected by the same issues

6. Re-run the models, and repeat the steps above until there are none that have a large Cook's distance

5.3 Stage 3: Investigate Data Points with Large Residuals

The aim is not to remove all data points which have large residuals but to use the residual to identify data points which might have other quality issues that can be identified using the checks for internal/external consistency described. This stage improves data quality, which refines the model estimates.

1. Run models
2. Rank the data points according to their residuals
3. Begin by investigating the largest negative residuals using the checks for internal and external consistency to identify any of the sources of error described above
4. Investigate all other data points reported by the company and assess whether they are affected by the same issues
5. Where appropriate, apply data reliability flags to the values and remove from the model sample
6. Re-run the models after removing invalid observations. Note that after stage 2 the models should remain stable.

5.4 Stage 4: Stopping the process

After stage 2, the model results should remain reasonably constant as more observations are dropped. As the analyst investigates the larger residuals, flagging them where indications of misreporting have been found, and moving through each one in turn, a point comes where there is little reason to remove any of the remaining data points which have large residuals. These remaining large residuals could be the natural extremes of what the statistical models consider likely. In the cases where there is no obvious evidence of misreporting then it would not be appropriate to remove the observation, a large residual alone is not a good enough reason to flag a data point.

Investigating the data points with large residuals in sufficient detail is a time-consuming task. Performing all the checks and trying to diagnose the problem can, at times, take up to an hour to fully assess just one response. As CDP has a limited number of analysts working on this project it is not possible to investigate the responses from every company to this level of detail (over 5,000 companies are included in the dataset). Prioritisation is necessary:

- First, the data points that fail the basic checks for internal consistency are removed
- Then the models are cleaned by removing the largest outliers that are distorting the estimates
- Once the models are reasonably robust, they can be used to identify the less obvious outliers. Each one of these is then investigated to see if it has been misreported.

There may still be misreported data in the final dataset that happens to fall within the range expected by the models. These false negatives are difficult to avoid without increasing the level of investigation and engagement with companies.

Appendix 1: CDP's Programs

- **Investor Program:** for this program CDP reaches out to the world's largest listed companies, asking them to respond to its information requests on behalf of institutional investors.

- **Supply Chain Program:** Supply Chain Members are companies that want to understand the emissions from their suppliers; CDP collects and analyses this data on their behalf. Respondents to the Supply Chain Program include privately owned companies which only disclose the information on the condition that CDP does not share it with anyone other than the Supply Chain Member which requested it.

Appendix 2: CDP Activity Classification System

Rationale

The CDP Activity Classification System (CDP ACS) has been developed to classify companies according to their environmental impacts across CDP's three themes: Water, Forests and Climate Change. This enables CDP to focus its work by categorizing companies in a way that lends itself to discussing environmental issues. Existing classification systems have a tendency to categorize companies looking at correlated investment risk, for example: *which companies would be worst affected by a crash in commercial property prices?* The answer would include builders, engineering consultancies, architects and REITS. Many existing classification systems would group these companies together for that reason. CDP has often struggled with using these other classification systems because these companies all have very different environmental impacts. The CDP classification system makes the distinction between these companies based on their environmental impacts.

Classification Hierarchy

CDP has defined a three-tiered classification hierarchy:

- **Activities:** environmental impacts can be traced to a business activity, and so the most granular level of the CDP classification system is called the activity. After examining revenue breakdown data from Bloomberg, CDP has defined 208 activities

- **Activity Groups:** CDP grouped the 208 activities into 60 activity groups.

- **Industries:** finally, CDP has grouped the 60 sectors into 13 industries.

Methodology

- CDP categorizes companies using revenue breakdown information from company filings, obtained from Bloomberg.
- Four impact categories have been defined for each of the three programs (12 in total), and these have been scored as medium/high/low etc.
- Activities with similar scores across the 12 impact categories have been grouped together into sectors and these sectors have been grouped into industries.

Climate Change Hybrid Classification System

The CDP classification has been a leap forward compared with the mix of GICS and BICS that was used in the 2015 modelling work. The full CDP classification system has been adjusted to improve the statistical models used to check reported data and estimate for the emissions of nondisclosers. These models are built using data reported to CDP by companies that are classified under the CDP classification system and rely heavily on having companies classified correctly. Some sectors have different response rates for different CDP questionnaires, because this work focuses on Climate Change the CDP sector classification system has been adapted to focus on the Climate Change theme. Some CDP activities were defined with Forest impacts in mind; for example, there are 19 different categories of Agricultural Producers. In many of these activity groups the climate change response rate is low, or even zero. Whereas the response rates for the Forests questionnaire are much higher for companies in the Agricultural Producers sector. At first these 19 sparsely populated activity groups resulted in poor model estimates for this sector, now, they have been re-grouped into just three and this has improved the reliability of the model.

The result is the Climate Change Hybrid Sector Classification system, an adapted version of the CDP Classification System that has been tailored to produce the most reliable models.