CDP Full GHG Emissions Dataset

Technical Annex II: Statistical Framework

© 2025 Google

Contents

1 Introduction 1

2 Modelling Assumptions 1

2.1 Revenue as a proxy for production 2

2.2 Sector & activity information 2

2.3 Variation between countries is larger than variation between activities 3

2.4 Each Scope is treated in isolation of the other 3

2.5 Consistent reporting 4

3 Modelling Process 4

3.1 Data structure 4

3.2 Model selection 5

3.3 Applying the models to the data 6

3.4 Limitations and sources of bias 6

Appendix 1 – Climate Change Hybrid Classification example 7

1 Introduction

The Full GHG Emissions Dataset builds on self-reported data disclosed through CDP to provide a comparable, comprehensive and consistent corporate GHG emissions and energy use dataset.

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 2024
- Technical Annex I: Data Cleaning Approach
- Technical Annex II: Statistical Framework
- Technical Annex III: Scope 3 Overview and Modelling

In this document, a statistical framework is defined for comparing the reported emissions data of companies on a like-for-like basis across all sectors. This framework allows us to establish what should be considered normal for any company by comparing it with other similar companies.

This is achieved by building statistical models, which serves two purposes: (1) allowing reported data to be reviewed so that outliers can be identified and investigated, and (2) producing estimates of company emissions when this information is not reported.

The statistical models described in this document 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 energy* from a third party. The CDP Climate Change questionnaire accounts for four types of purchased energy: steam, heat, electricity and cooling (SHEC).

The modelling approach for Scope 3 emissions is discussed in more detail in *Technical Annex III: Scope 3 Overview and Modelling.*

2 Modelling Assumptions

A statistical model can be thought of as framework used to describe a set of observations. Discrepancies between the observed data and the model can be reduced by making the model framework more complicated, but there are practical and mathematical limitations. These include:

- Data availability: This framework must be applicable to companies from all sectors and so any parameters (e.g. company revenue) must be available for all companies in the sample.
- Sample size: A limited number of companies report emissions figures, restricting the amount of information in the model.
- Over-fitting: Arbitrarily increasing the number of model parameters can result in an overfitted model, which detects not only the underlying trend in the data, but also the noise in the sample. Thus, an overfitted model may have a high predictive accuracy on the training data, but low predictive accuracy for the real-world population.

- Multi-collinearity: Adding model parameters that are similar to those already included may give the illusion of a more sophisticated model, whereas in fact they may not provide any additional understanding of the emissions of different companies. In these cases, a simpler model is preferable.

Building one emissions model that fits all companies in all sectors of the global economy is a very complex problem. To make the problem more manageable, generalising assumptions are made. However, these introduce sources of inaccuracy in the model because they ignore certain complexities. With this in mind, each assumption should be assessed in terms of its *cost* due to loss of accuracy, the *benefit* it brings by simplifying the problem, and how *reasonable* it seems when applied to all companies.

Where the models differ from the reported data, the root cause will be related to one of the following assumptions.

2.1 Revenue as a proxy for production

- *Revenue from an activity is directly proportional to production.* This implicitly assumes that each unit of production is sold at a constant price.
- *Revenue from an activity is directly proportional to emissions.* If the previous assumption is made, then we can also assume there is a constant revenue intensity *tCO2e/$* for each of a company's different activities. However, this is demonstrably untrue as the price of commodities varies daily whereas the energy and processes that produce them stay relatively constant. To mitigate the effect of these price fluctuations CDP treats each year's data independently of other years' data.
- *Zero Revenue = Zero Activity = Zero Emissions.* This assumption is implied by the assumptions above, but in practice, it has significant statistical implications and so warrants a separate discussion. There are many cases where companies have had non-positive revenue. For example, if a finance company reported negative earnings due to trading losses, this would imply that their emissions would have had to have been negative. They are operating under exceptional circumstances where their revenue model has been knocked by market forces, and so they breach the assumption that revenue is proportional to emissions.

2.2 Sector & activity information

CDP have developed an activity classification system that groups companies according to the environmental impacts of their activities across three themes: Water, Forests and Climate Change. It is the activities of a company that have impacts on the environment, and so the most granular tier of the CDP classification tree is called the 'CDP Activity'. The 214 Activities are combined to form 62 'Activity Groups', which in turn are grouped into 13 'Industries'. Details of the CDP activity classification system can be found in the Appendix of the *Technical Annex I: Data Cleaning Approach*.

This classification system provides a structure for defining the activities discussed in the previous section on revenue. Whereas previously each company only had one sector or sub-industry classification, companies can now have more than one CDP Activity using the CDP classification system.

Grouping company activities together into any classification system means that there are several implicit assumptions being made, as summarised below:

- All companies engaged in each CDP Activity produce similar products.
- All companies engaged in each CDP Activity sell these products at a similar price.
- All companies engaged in each CDP Activity produce these products in a similar way.

An adapted version of the CDP Activity Classification System has been used in this dataset, focusing exclusively on the Climate Change related impacts, rather than Climate Change, Water and Forests. This adapted classification system was developed to optimise the amount of differentiation between company activities whilst adhering to the constraints and limitations discussed above. The adapted CDP classification system is called the Climate Change Hybrid because it is a hybrid of different levels of the CDP classification system. The Climate Change Hybrid has grouped together some of the smaller CDP activities together based on the data available for this project. An explanation can be found in the Appendix.

2.3 Variation between countries is larger than variation between activities

Both regional and activity revenue breakdowns are used in our modelling. To use both variables in a single model would require sectoral revenue breakdowns for each region in which a company operates. This level of granularity is not present in the data available to us. Due to this limitation, the models can use *either* the regional *or* the divisional revenue breakdown, but not both.

While there are usually regional differences in the Scope 1 emissions intensity of any given activity due to differences in regulations, technology etc., these are assumed to be negligible in comparison with the variation in emissions factors between activities. For example, there is a discernible difference in Scope 1 emissions factors between the US steel industries and other countries on an average tonne-for-tonne basis. However, this variation in Scope 1 emissions intensity between global steel industries is smaller than variation of different industries within the US.

This assumption holds in the Scope 1 emissions for both fuel use and SHEC models. Location-based Scope 2 emissions are treated differently, because the regional variation of emissions intensities within any given sector will be large due to the differences in grid emissions factors.

2.4 Each Scope is treated in isolation of the other

We make the assumption that each scope is independent from one another. The relationship between Scope 1 and Scope 2 is more complex and companies may choose to purchase energy from a third party rather than generate it themselves. This follows from the assumption that *companies in each CDP activity group manufacture their products in the same way*. Three companies may manufacture identical products using the same equipment, but they may power the electric equipment by different means. For example, one company may use a diesel generator to power the electric equipment, another may use electrical equipment powered from the mains, and a third may use rooftop solar to generate their electricity. Clearly, these three companies would all have very different emissions profiles.

2.5 Consistent reporting

This framework does not explicitly account for the variability between reporting standards, reporting boundaries, and calculation methodologies. That does not mean that these factors are being ignored, but rather that these issues are dealt with on a case-by-case basis by CDP's analysts. Many companies are not able to collect emissions data from across their whole organisation because they are still refining their data collection and accounting methodologies. For emissions data to be comparable, one must assume that companies' reported data covers the full scope of their operations.

For some companies where these differences are most pronounced it has been necessary to exclude them from the modelling sample because they do not fit with the basic assumption that all companies are using similar accounting practises to measure their emissions. These issues are very complex, and the approach for dealing with them is discussed in *Technical Annex I: Data Cleaning Approach*.

3 Modelling Process

3.1 Data structure

Before any statistical analysis can begin, the data used for comparing companies must be collected and organised. To maximise the amount of available information in the model, data from private companies reporting through the CDP Supply Chain program has been included.

By exploring the relationship between company emissions and revenue, several observations are made:

- The relationship between revenue and emissions is positive
- Both emissions and revenue vary across many orders of magnitude. Most data points clustered at lower magnitude, with several notable outliers at higher magnitudes.
- There is a large variation in intensity (emissions per unit revenue). This often depends on the activity in question: e.g. power generation has a particularly high intensity, whereas the services industry has a particularly low intensity.
- The data is heteroscedastic. This means that variance of emissions increases with revenue, i.e. the greater the revenue, the greater the range in reported emissions.

This latter property poses a challenge to the application of standard regression analysis, which typically assumes that the modelling errors are uniform across the distribution (i.e.

homoscedastic). The choice of statistical model must account for this, without compromising any of the other assumptions made.

3.2 Model selection

This project uses regression analysis to model company emissions and energy consumption across all sectors. Regression modelling involves relating a dependent variable (the value we want to estimate), to one or more independent variables, which are commonly referred to as ‘predictors’.

In our case, the dependent variable is either Scope 1 emissions, fuel, or SHEC consumption. Following from the assumptions made earlier, the independent variables are company activity (as defined by the CDP Activity Classification System), and activity-revenue (the revenue of a company from that particular activity).

As explained previously, this choice of model parameters is a compromise between precision and practicality. While adding more predictor variables may increase the accuracy of the model, the additional data points would need to be accessible for all companies in the sample, as well as any companies for which an estimate is required. Given that our dataset includes over 14,500 companies across all sectors and regions, limiting the number of predictor variables has the advantage of increasing the scalability of the model.

In the previous section we established that the dataset exhibits heteroscedasticity (i.e. that the spread of emissions tends to increase with revenue). Another way of putting this is that the model residuals are not normally distributed. To account for this, two regression modelling methods are considered:

- *Generalised Linear Model:* This is a flexible regression modelling framework that allows for dependent variables with residuals that are not normally distributed.
- *Logarithmic regression:* This method applies a transformation to the dataset with a logarithmic function. For our dataset, this has the effect of removing the skew from the distribution of residuals.

Both modelling techniques account for the heteroscedasticity of the dataset. The key difference that logarithmic regression applies a transformation to the data, whereas the GLM does not. Since we assume that company emissions are directly proportional to revenue, applying the logarithmic transformation undermines this assumption. Therefore, the GLM is the primary regression modelling framework used in this project.

There are many varieties of GLMs, and their applicability depends on the properties of the dataset in question. Of the range of candidate GLMs that were considered, the Gamma GLM yielded the most positive results with regards to its quality of fit.

3.3 Applying the models to the data

Having defined the modelling framework, it is now necessary to return to our initial aim, which is to produce estimates for the following variables: Scope 1 emissions, Scope 2 emissions (location based), fuel consumption and SHEC consumption.

The regression models described in the previous section are suitable to model all these variables, apart from Scope 2 emissions. This is because this variable has a large regional variation due to the different grid emissions factors. To estimate for location-based Scope 2 emissions, CDP multiplies the IEA national level grid emissions factors by the SHEC estimates. Where a company provides a regional revenue breakdown in their company filings, a company specific revenue weighted emissions factor is used to calculate the Location-based Scope 2 emissions.

Before estimates are produced, CDP's team of analysts first cleaned the data, as documented in *Technical Annex I: Data Cleaning Approach.* Data points are not removed from the model training sample if the company has reported inconsistent data. In addition, values that have a disproportionate influence on the model estimates are investigated and removed if appropriate.

3.4 Limitations and sources of bias

The primary limitation of these models is that several simplifying assumptions are made to maximise their scalability, as discussed in previous sections. A further limitation is statistical bias, which occurs when the observed sample is not representative of the population. In this project, the population is the entire corporate universe, and the observed sample comprises of the companies reporting to CDP, excluding any data judged to have been misreported. Where the sample differs from the population, it is most likely due to the following sources of bias:

- Reporting bias: Companies disclosing their emissions are assumed to have lower emissions intensities than those that do not disclose their emissions. This is because companies who are engaged in GHG reporting are more likely to take measures that reduce their environmental footprint. The resulting bias would result in underestimates for nonreporting companies.
- *Cleaning bias:* The context of an organisation is important when considering the validity of their reported data which adds an extra layer of complexity. As a result, the analysts cleaning the data have had to use subjective judgements to treat outliers. *Technical Annex I: Data Cleaning Approach* provides a summary of the basis for identifying and flagging potentially misreported data.
- Model Bias: The y-intercept is not fixed to zero in the statistical models, which can result in a small positive bias to the estimates of smaller companies. However, for the purposes of this project, this is deemed to be negligible.

Appendix 1 – Climate Change Hybrid Classification example

The issue of sample size is the primary reason for creating the Climate Change Hybrid classification system. For example, in the Crop Farming Activity Group there is a separate activity for Soybean Farming, Grain Farming, Fruit Farming and Cotton Farming because of their distinct water consumption patterns. The sample sizes for these categories are small, which reduces the robustness of the statistical model.

Consequently, several CDP Activities have been grouped together to form a new category in the Climate Change Hybrid. Those CDP activities are:

- Animal feed
- Cocoa farming
- Coffee
- Crop farming
- Fruit farming
- Grain farming
- Palm oil & oilseed farming
- Rice farming
- Soybean farming
- Sugarcane farming
- Tea
- Vegetable farming
- Cotton farming
- Rubber plantation

Similar compromises have been made across all sectors in order to achieve a Climate Change Hybrid sector classification system that has distinct activities with adequately sized samples. The resultant Climate Change Hybrid sector classification system has 89 distinct activities, reduced from the 214 distinct activities that make up the full CDP sector classification tree.